

**ANALISIS SENTIMEN KOMENTAR MASYARAKAT
TERHADAP KEBIJAKAN PEMERINTAH TENTANG
SISTEM ZONASI SEKOLAH MENGGUNAKAN
ALGORITMA *K-MEANS* DAN
ALGORITMA *LEVENSTHEIN DISTANCE***

Skripsi



Oleh :

Muhammad Haris Al Farisi

11140910000056

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH

JAKARTA

2019 M / 1440 H

**ANALISIS SENTIMEN KOMENTAR MASYARAKAT
TERHADAP KEBIJAKAN PEMERINTAH TENTANG
SISTEM ZONASI SEKOLAH MENGGUNAKAN
ALGORITMA *K-MEANS* DAN
ALGORITMA *LEVENSTHEIN DISTANCE***

Skripsi

Diajukan sebagai salah satu syarat untuk memperoleh gelar S1

Sarjana Komputer (S.Kom)



Oleh :

Muhammad Haris Al Farisi

11140910000056

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH

JAKARTA

2019 M / 1440 H

LEMBAR PERSETUJUAN
ANALISIS SENTIMEN KOMENTAR MASYARAKAT
TERHADAP KEBIJAKAN PEMERINTAH TENTANG SISTEM
ZONASI SEKOLAH MENGGUNAKAN
ALGORITMA *K-MEANS* DAN
ALGORITMA *LEVENSTHEIN DISTANCE*

Skripsi

Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Komputer Pada
Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta

Oleh:

Muhammad Haris Al Farisi

11140910000056

Menyetujui,

Pembimbing I

Pembimbing II



Arini, MT

NIP. 19760131 200901 2 001



Luh Kesuma Wardhani, MT

NIP. 19780424 200801 2 022

Mengetahui,

Ketua Program Studi Teknik Informatika



Arini, MT

NIP. 19760131 200901 2 001

PENGESAHAN UJIAN

Skripsi berjudul “Analisis Sentimen Komentar Masyarakat Terhadap Kebijakan Pemerintah Tentang Sistem Zonasi Sekolah Menggunakan Algoritma *K-Means* Dan Algoritma *Levensthein Distance*” yang ditulis oleh M. Haris Alfarisi, NIM 11140910000056 telah diuji dan dinyatakan lulus dalam sidang *munaqosyah* Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta pada Februari 2019. Skripsi ini telah diterima sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) pada Program Studi Teknik Informatika.

Jakarta, Januari 2019

Tim Penguji

Penguji I,

Khodijah Hulliyah, M.Si
NIP. 19730402 200112 2 001

Penguji II,

Nurul Faizah Rozy, MTI
NIDN. 2009027202

Tim Pembimbing

Pembimbing I,

Arini, MT
NIP. 19760131 200901 2 001

Pembimbing II,

Luh Kesuma Wardhani, MT
NIP. 19780424 200801 2 022

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Prof. Dr. Lily Surraya Eka Putri,
M.Env.Stud
NIP. 196904042005012005

Ketua Program Studi Teknik Informatika

Arini, MT
NIP. 197601312009012001

PERNYATAAN ORISINALITAS

DENGAN INI SAYA MENYATAKAN BAHWA SKRIPSI INI BENAR-BENAR HASIL KARYA SENDIRI YANG BELUM PERNAH DIAJUKAN SEBAGAI SKRIPSI ATAU KARYA ILMIAH PADA PERGURUAN TINGGI ATAU LEMBAGA MANAPUN.

Jakarta, Februari 2019



Muhammad Haris Al Farisi



PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai civitas akademik UIN Syarif Hidayatullah Jakarta, saya yang bertanda tangan di bawah ini:

Nama : Muhammad Haris Al Farisi

NPM : 11140910000056

Program Studi : Teknik Informatika

Departemen : Teknik Informatika

Fakultas : Sains dan Teknologi

Jenis Karya : Skripsi

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Islam Negeri Syarif Hidayatullah Jakarta Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty Free Right*) atas karya ilmiah yang berjudul:

**ANALISIS SENTIMEN KOMENTAR MASYARAKAT TERHADAP
KEBIJAKAN PEMERINTAH TENTANG SISTEM ZONASI SEKOLAH
MENGUNAKAN ALGORITMA *K-MEANS* DAN
ALGORITMA *LEVENSTHEIN DISTANCE***

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Islam Negeri Syarif Hidayatullah Jakarta berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilih Hak Cipta. Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, Februari 2019

Yang menyatakan



(Muhammad Haris Al Farisi)

KATA PENGANTAR

Assalamu'alaikum Wr. Wb.

Puji syukur senantiasa dipanjatkan kehadirat Allah SWT yang telah melimpahkan rahmat, hidayah serta nikmat-Nya sehingga penyusunan skripsi ini dapat diselesaikan. *Sholawat* dan salam senantiasa dihaturkan kepada junjungan kita baginda Nabi Muhammad SAW beserta keluarganya, para sahabatnya serta umatnya hingga akhir zaman. Penulisan skripsi ini mengambil tema dengan judul:

ANALISIS SENTIMEN KOMENTAR MASYARAKAT TERHADAP KEBIJAKAN PEMERINTAH TENTANG SISTEM ZONASI SEKOLAH MENGUNAKAN ALGORITMA *K-MEANS* DAN ALGORITMA *LEVENSTHEIN DISTANCE*

Penyusunan skripsi ini adalah salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) pada program studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta. Adapun bahan penulisan skripsi ini adalah berdasarkan hasil penelitian, pengembangan aplikasi, kuesioner dan beberapa sumber literatur.

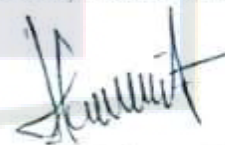
Dalam penyusunan skripsi ini, telah banyak bimbingan dan bantuan yang didapatkan dari berbagai pihak sehingga skripsi ini dapat berjalan dengan lancar. Oleh karena itu, penulis ingin mengucapkan banyak terima kasih kepada :

1. Bapak Dr. Agus Salim, M.Si selaku dekan Fakultas Sains dan Teknologi.
2. Ibu Arini, MT. selaku Ketua Program Studi Teknik Informatika.
3. Ibu Arini, MT. dan Ibu Luh Kesuma Wardhani, MT. selaku Dosen Pembimbing I dan II yang senantiasa meluangkan waktu dan memberikan bimbingan, bantuan, semangat dan motivasi dalam menyelesaikan skripsi ini.
4. Seluruh Dosen, Staf Karyawan Fakultas Sains dan Teknologi, khususnya Program Studi Teknik Informatika yang telah memberikan bantuan dan kerja sama dari awal perkuliahan.

5. Orang tua tercinta, Umi dan Abi yang senantiasa memberikan dukungan moril dan materil. Tiada tutur kata selain terima kasih kepada kalian. Terima kasih Umi dan Abi, Ais sayang sama Umi dan Abi.
6. Adik – adik tersayang, Fahmi, Fathiyah. Farihah, dan Fathan yang juga senantiasa memberi dukungan doa, semangat, dan motivasi dalam menyelesaikan skripsi ini.
7. Kepada teman seperjuangan Teknik Informatika angkatan 2014 dan teman – teman KKN Komika, yang sudah membantu penulis dalam menyelesaikan skripsi ini, terima kasih atas dukungannya. Semoga kita bisa lebih baik lagi dan sukses di masa yang akan datang.
8. Seluruh pihak yang secara langsung maupun tidak langsung membantu penulis dalam menyelesaikan skripsi ini.

Akhir kata, penulis menyadari bahwa dalam penyajian skripsi ini masih jauh dari sempurna. Apabila ada kebenaran dari penulisan ini maka kebenaran tersebut datang dari Allah, tetapi apabila ada kesalahan dalam penulisan ini maka kesalahan ini berasal dari penulis. Semoga skripsi ini membawa manfaat bagi pengembangan ilmu. Penulis berharap Allah SWT berkenan membalas segala kebaikan semua pihak yang telah membantu dan meridhoi segala usaha kita.

Jakarta, Februari 2019



Muhammad Haris Al Farisi

11140910000056

Nama : Muhammad Haris Al Farisi
Program Studi : Teknik Informatika
Judul : Analisis Sentimen Komentar Masyarakat Terhadap Kebijakan Pemerintah Tentang Sistem Zonasi Sekolah Menggunakan Algoritma *K-Means* dan Algoritma *Levenshtein Distance*

ABSTRAK

Peranan pendidikan sangat besar pengaruhnya dalam menggapai kemajuan bagi setiap bangsa dan negara. Kemendikbud Republik Indonesia, menetapkan peraturan baru yaitu Permendikbud nomor 14 tahun 2018 tentang Penerimaan Peserta Didik Baru. Namun di dalam Permendikbud tersebut terdapat kebijakan yang menimbulkan pro dan kontra pada masyarakat yaitu dengan diterapkannya sistem zonasi. Pada penelitian ini dilakukan analisis sentimen dari komentar masyarakat terhadap kebijakan tersebut. Data yang dianalisis yaitu data yang diambil dari Facebook Kemendikbud dan YouTube *channel* CNN Indonesia sebanyak 200 komentar. Penelitian ini menggunakan algoritma *K-Means* dengan nilai $k=2$ untuk menentukan sentimen akhir yaitu positif dan negatif dan algoritma *Levenshtein Distance* untuk normalisasi kata. Hasil yang didapatkan dari sentimen masyarakat terhadap kebijakan ini yaitu lebih banyak yang bersentimen negatif daripada yang positif. Hasil dari tingkat akurasi yang didapatkan dari penggunaan algoritma *K-Means* saja yaitu 84% dengan nilai akurasi yang lebih rendah dibandingkan dengan kombinasi algoritma diatas yaitu 90%. Saran untuk penelitian selanjutnya sebaiknya menggunakan data yang lebih banyak lagi, dan menggunakan teknik perhitungan akurasi *k-fold cross validation* sebagai uji coba selanjutnya.

Kata Kunci : Sentimen analisis, zonasi sekolah, *k-means*, *levenshtein distance*, *k-fold cross validation*.
Jumlah Pustaka : 4 buku, 2 *e-book*, dan 22 jurnal.
Jumlah Halaman : 135 halaman

Name : Muhammad Haris Al Farisi
Program Studi : Teknik Informatika
Title : Sentiment Analysis of Community Comments on Government Policy About School Zoning Systems Using K-Means Algorithm and Levenshtein Distance Algorithm

ABSTRACT

The role of education is very influential in achieving progress for every nation and country. The Ministry of Education and Culture of the Republic of Indonesia has set a new regulation namely Permendikbud number 14 of 2018 concerning New Student Admissions. However, in the Permendikbud there is a policy that raises the pros and cons of the community, namely the implementation of the zoning system. In this study an analysis of the sentiments of community comments on the policy was carried out. The data analyzed are data taken from Facebook Ministry of Education and Culture and YouTube CNN Indonesia channel as many as 200 comments. This study uses the K-Means algorithm with a value of $k = 2$ to determine the final sentiment that is positive and negative and the Levenshtein Distance algorithm for word normalization. The results obtained from public sentiment towards this policy are more negative than positive ones. The results of the level of accuracy obtained from the use of the K-Means algorithm are only 84% with a lower accuracy value compared to the combination of the above algorithms which is 90%. Suggestions for further research should use more data, and use the technique of calculating accuracy of k-fold-cross validation as the next trial.

Keywords : Analysis sentiment, school zone, k-means, levensthein distance, k-fold cross validation.

Bibliography : 5 books dan e-book, 22 journals.

Number of Pages : 135 pages

DAFTAR ISI

LEMBAR PERSETUJUAN.....	2
PENGESAHAN UJIAN	iii
PERNYATAAN ORISINALITAS	iv
PERNYATAAN PERSETUJUAN PUBLIKASI	v
KATA PENGANTAR	vi
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR GAMBAR	xiv
DAFTAR TABEL	xv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	6
1.3 Batasan Masalah	6
1.4 Tujuan Masalah	7
1.5 Manfaat Penelitian	7
1.5.1 Bagi Penulis	7
1.5.2 Bagi Universitas	8
1.5.3 Bagi Pemerintah	8
1.6 Metodologi Penelitian	8
1.6.1 Metode Pengumpulan Data	8
1.6.2 Metode Implementasi	8
1.7 Sistematika Penulisan	9
BAB 2 LANDASAN TEORI	11
2.1 Analisis Sentimen	11
2.1.1 Level Analisis Sentimen	11
2.2 Sistem Zonasi Sekolah	12
2.3 <i>Clustering</i>	12
2.4 Algoritma	12

2.5	<i>Text Mining</i>	13
2.6	<i>Pre-processing</i>	13
2.6.1	<i>Case Folding</i>	14
2.6.2	<i>Tokenizing</i>	14
2.6.3	<i>Filtering/Stopword</i>	14
2.6.4	<i>Stemming</i>	14
2.7	Algoritma Nazief dan Adriani	15
2.7.1	Tahapan Algoritma Nazief dan Adriani	15
2.7.2	Alasan Menggunakan Algoritma Nazief dan Adriani.....	16
2.8	Algoritma <i>K-Means Clustering</i>	17
2.8.1	Penjelasan Algoritma <i>K-Means Clustering</i>	17
2.8.2	Tahapan Algoritma <i>K-Means Clustering</i>	17
2.9	Algoritma <i>Levenshtein Distance</i>	18
2.8.1	Penjelasan Algoritma <i>Levenshtein Distance</i>	18
2.8.2	Tahapan Algoritma <i>Levenshtein Distance</i>	19
2.10	Algoritma TF-IDF.....	21
2.11	<i>Confusion Matrix</i>	23
2.12	PHP	23
2.13	MySQL	23
2.14	<i>R-Programming</i>	24
2.15	Metode Pengumpulan Data.....	24
2.15.1	Teknik Wawancara.....	24
2.15.2	Teknik Observasi	25
2.15.3	Teknik Kuisisioner.....	26
2.16	Metode Simulasi	27
2.16.1	Formulasi Masalah (<i>Problem Formulation</i>)	27
2.16.2	Model Pengkonsepian (<i>Conceptual Model</i>).....	27
2.16.3	Data Masukan Keluaran (<i>Input Output Data</i>).....	28
2.16.4	Pemodelan (<i>Modelling</i>)	28
2.16.5	Simulasi (<i>Simulation</i>).....	29
2.16.6	Verifikasi dan Validasi (<i>Verification and Validation</i>).....	29
2.16.7	Experimentasi (<i>Experimentation</i>)	29

2.16.8	Analisa Keluaran (<i>Output Analysis</i>).....	29
2.17	Studi Literatur Sejenis	30
BAB 3 METODOLOGI PENELITIAN.....		34
3.1	Metode Pengumpulan Data	34
3.2.1	Studi Lapangan.....	34
3.2.2	Studi Pustaka.....	34
3.2	Metode Simulasi.....	35
3.2.1	Formulasi Masalah (<i>Problem Formulation</i>)	35
3.2.2	Model Pengkonsepian (<i>Conceptual Model</i>).....	35
3.2.3	Data Masukan/Keluaran (<i>Input/Output Data</i>)	36
3.2.4	Pemodelan (<i>Modelling</i>).....	36
3.2.5	Simulasi (<i>Simulation</i>)	37
3.2.6	Verifikasi dan Validasi (<i>Verification and Validation</i>)	37
3.2.7	Eksperimentasi (<i>Experimentation</i>).....	37
3.2.8	Analisis Keluaran (<i>Output Analysis</i>).....	38
3.3	Kerangka Berfikir.....	38
BAB 4 IMPLEMENTASI, SIMULASI, DAN EKSPERIMEN		40
4.1	Formulasi Masalah (<i>Problem Formulation</i>).....	40
4.2	Model Pengkonsepian (<i>Conceptual Model</i>)	40
4.2.1	<i>Conceptual Model</i> Pada <i>Text Mining</i>	40
4.2.2	<i>Conceptual Model</i> Sentimen Pada Data Latih	46
4.2.3	<i>Conceptual Model</i> Sentimen dengan Algoritma <i>K-Means</i>	48
4.2.4	<i>Conceptual Model</i> Sentimen Algoritma <i>K-Means</i> dengan Bantuan Algoritma <i>Levenshtein Distance</i>	51
4.3	Data Masukan/Keluaran (<i>Input / Output Data</i>).....	53
4.4	Pemodelan (<i>Modelling</i>)	54
4.4.1	Konstruksi Sentimen dengan Algoritma <i>K-Means</i>	54
4.4.2	Konstruksi Sentimen Algoritma <i>K-Means</i> dengan Bantuan Algoritma <i>Levenshtein Distance</i>	79
4.5	Simulasi (<i>Simulation</i>)	104
4.5.1	Tahap Pengujian Data Uji	105
4.6	Verifikasi dan Validasi (<i>Verification and Validation</i>)	106
4.7	Eksperimentasi (<i>Experimentation</i>)	107

4.8	Analisis Keluaran (<i>Output Analysis</i>)	107
BAB 5 HASIL DAN PEMBAHASAN.....		108
5.1	Verifikasi dan Validasi (<i>Verification and Validation</i>)	108
5.2	Eksperimentasi (<i>Experimentation</i>)	109
5.3	Analisis Keluaran (<i>Output Analysis</i>)	109
5.3.1	Hasil Sentimen Algoritma <i>K-Means</i> dan Kombinasi Algoritma <i>K-Means</i> dan Algoritma <i>Levensthein Distance</i>	109
5.3.2	Analisis Hasil Akurasi Algoritma <i>K-Means</i>	112
5.3.3	Analisis Hasil Akurasi Kombinasi Kombinasi Algoritma <i>K-Means</i> dan Algoritma <i>Levensthein Distance</i> dan Sentimen Manual	113
5.3.4	Analisis Hasil Peningkatan Akurasi Kombinasi Algoritma <i>K-Means</i> dan Algoritma <i>Levensthein Distance</i> dibandingkan dengan Algoritma <i>K-Means</i> saja.....	114
BAB 6 PENUTUP		116
6.1	Kesimpulan.....	116
6.2	Saran	116
DAFTAR PUSTAKA		118
LAMPIRAN		121

DAFTAR GAMBAR

Gambar 1.1 Presentase Penggunaan Aplikasi Media Sosial di Indonesia.....	3
Gambar 2.1 Tahap <i>Pre-processing</i>	13
Gambar 3.1 Kerangka Berfikir	39
Gambar 4.1 Flowchart Tahapan Pre-Processing	41
Gambar 4.2 <i>Flowchart Case Folding</i>	42
Gambar 4.3 <i>Flowchart Tokenization</i>	43
Gambar 4.4 <i>Flowchart Levensthein Distance</i>	44
Gambar 4.5 <i>Flowchart Stopword Removal</i>	45
Gambar 4.6 <i>Flowchart Stemming</i> Algoritma Nazief dan Adriani.....	46
Gambar 4.7 Flowchart penentuan sentimen data latih	47
Gambar 4.8 <i>Flowchart</i> Proses Sentimen Skenario 1	49
Gambar 4.9 <i>Flowchart</i> Proses Algoritma <i>K-Means</i>	50
Gambar 4.10 <i>Flowchart</i> Proses Sentimen Skenario 2.....	52
Gambar 4.11 Hasil Akurasi dari Aplikasi Skenario 1	106
Gambar 4.12 Hasil Akurasi dari Aplikasi Skenario 2	106
Gambar 5.1 Grafik Peningkatan Akurasi Pada Skenario 2	114



DAFTAR TABEL

Tabel 2.1 Tingkat Akurasi Algoritma-Algoritma <i>Stemming</i>	17
Tabel 2.2 Model <i>Confusion Matrix</i>	23
Tabel 2.3 Studi Literatur Sejenis	31
Tabel 4.1 Contoh Proses <i>Case Folding</i>	47
Tabel 4.2 Contoh Proses <i>Tokenization</i>	47
Tabel 4.3 Contoh Proses Normalisasi Kata	48
Tabel 4.4 Contoh Proses <i>Stopword Removal</i>	48
Tabel 4.5 Contoh Proses <i>Stemming</i>	48
Tabel 4.6 Dokumen data latih	54
Tabel 4.7 Penentuan sentimen data latih	55
Tabel 4.8 Hasil Proses <i>Case Folding</i>	57
Tabel 4.9 Hasil Proses <i>Filtering</i>	57
Tabel 4.10 Hasil Proses <i>Tokenization</i>	58
Tabel 4.11 Hasil Proses Normalisasi Kata	60
Tabel 4.12 Hasil Proses <i>Stemming</i>	63
Tabel 4.13 Hasil Proses <i>Filtering / Stopword</i>	65
Tabel 4.14 Hasil Perhitungan <i>IDF</i> Skenario 1	68
Tabel 4.15 Hasil Perhitungan bobot <i>TF-IDF</i> Skenario 2	70
Tabel 4.16 Perhitungan <i>Euclidean Distance cluster 1</i>	73
Tabel 4.17 Perhitungan <i>Euclidean Distance cluster 2</i>	73
Tabel 4.18 Perhitungan kembali <i>Euclidean Distance cluster 1</i>	75
Tabel 4.19 Perhitungan kembali <i>Euclidean Distance cluster 2</i>	76
Tabel 4.20 Dokumen data latih	79
Tabel 4.21 Penentuan sentimen data latih	80
Tabel 4.22 Hasil Proses <i>Case Folding</i>	81
Tabel 4.23 Hasil Proses <i>Filtering</i>	82
Tabel 4.24 Hasil Proses <i>Tokenization</i>	83
Tabel 4.25 Hasil Proses Normalisasi Kata	85
Tabel 4.26 Hasil Proses <i>Stemming</i>	87
Tabel 4.27 Hasil Proses <i>Filtering / Stopword</i>	90
Tabel 4.28 Perhitungan Matriks <i>Levensthein Distance 1</i>	92
Tabel 4.29 Perhitungan Matriks <i>Levensthein Distance 2</i>	92
Tabel 4.30 Hasil Perhitungan <i>IDF</i> Skenario 2	93
Tabel 4.31 Hasil Perhitungan bobot <i>TF-IDF</i> Skenario 2	95
Tabel 4.32 Perhitungan <i>Euclidean Distance cluster 1</i> Skenario 2	98
Tabel 4.33 Perhitungan <i>Euclidean Distance cluster 2</i> Skenario 2	99
Tabel 4.34 Perhitungan kembali <i>Euclidean Distance cluster 1</i>	101
Tabel 4.35 Perhitungan kembali <i>Euclidean Distance cluster 2</i>	101
Tabel 4.36 Faktor Simulasi Penelitian	105

Tabel 5.1 Hasil Sentimen dari Skenario Pada Data Uji.....	110
Tabel 5.2 Hasil Pengujian Skenario 1	112
Tabel 5.3 Hasil Pengujian Skenario 2	113



BAB 1

PENDAHULUAN

1.1 Latar Belakang

Peranan pendidikan sangat besar pengaruhnya dalam menggapai kemajuan bagi setiap bangsa dan negara. Dalam usaha mencapai tahap negara maju, pembentukan negara sangat bergantung dengan taraf pendidikan di suatu bangsa itu sendiri. Nilai pendidikan di suatu bangsa akan lenyap begitu saja jika bangsa tersebut lalai dan mudah terbawa arus globalisasi. Maksud lalai disini adalah apabila sebuah bangsa menganggap rendah pengaruh pendidikan bagi sebuah kemajuan bangsa tersebut. Indonesia salah satu negara yang menganggap bahwa pendidikan memiliki pengaruh besar dalam kemajuan negara. Dilihat dari Kementerian Pendidikan dan Kebudayaan (Kemendikbud) terus mengupayakan wajib belajar 12 tahun melalui pelaksanaan Program Indonesia Pintar (PIP).

Indonesia telah merumuskan definisi penting dari pendidikan itu sendiri. Menurut Kamus Besar Bahasa Indonesia (KBBI), pendidikan adalah proses pengubahan sikap dan tata laku seseorang atau kelompok orang dalam usaha mendewasakan manusia melalui upaya pengajaran dan pelatihan. Sedangkan pengertian dari ilmu pendidikan adalah pengetahuan tentang prinsip dan metode belajar, membimbing, dan mengawasi pelajaran.

Pada tanggal 2 Mei 2018, Menteri Pendidikan dan Kebudayaan Republik Indonesia, Muhadjir Effendy, menetapkan kebijakan baru yaitu Permendikbud nomor 14 tahun 2018 tentang Penerimaan Peserta Didik Baru. Dimana PPDB ini bertujuan untuk menjamin penerimaan peserta didik baru berjalan secara objektif, transparan, akuntabel, nondiskriminatif, dan berkeadilan dalam rangka mendorong peningkatan akses layanan pendidikan.

Sistem zonasi dalam kebijakan pemerintah tentang PPDB pada permendikbud no 14 tahun 2018 bagian IV, merupakan salah satu topik

perbincangan hangat di dalam media sosial pada bulan Juni 2018. Lebih dari 200 komentar masyarakat yang menyampaikan respon ataupun pendapat mereka tentang hal tersebut. Analisis sentimen atas pendapat-pendapat masyarakat tersebut dapat diteliti untuk mengetahui seberapa besar presentase yang bersentimen positif dan bersentimen negatif dari kebijakan ini melalui komentar-komentar yang telah dikirim ke sosial-media. Analisis sentimen (*sentiment analysis*) atau *opinion mining* sendiri merupakan topik riset yang penting dan sedang marak dilakukan saat ini. Analisis sentimen merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Terdapat beberapa penelitian sebelumnya yang berkaitan dengan analisis sentimen dengan objek yang berbeda-beda.

Dari kebijakan baru ini timbulah pro dan kontra diantara masyarakat yang merasakannya. Sebanding lurus dengan hasil wawancara pada tanggal 24 Oktober 2018 kepada ketua sub bagian Humas yaitu Any Sayekti yang berpendapat banyak pro dan kontra di dalam kebijakan baru ini, hal itupun sudah biasa. Seiring berjalannya waktu masyarakat akan menerima kebijakan baru ini. Pada dasarnya pemerintah membuat kebijakan baru ini demi kepentingan bersama. Tujuan lain dari yang sudah ditulis sebelumnya adalah guna untuk penyetaraan antara sekolah-sekolah negeri di Indonesia. Nanti tidak ada lagi yang namanya sekolah *favorite* ataupun sekolah unggulan seperti yang kita rasakan sebelumnya. Pemerintah ingin semua sekolah mempunyai kompetensi yang adil dan sama rata atau dengan kata lain tidak ada diskriminasi.

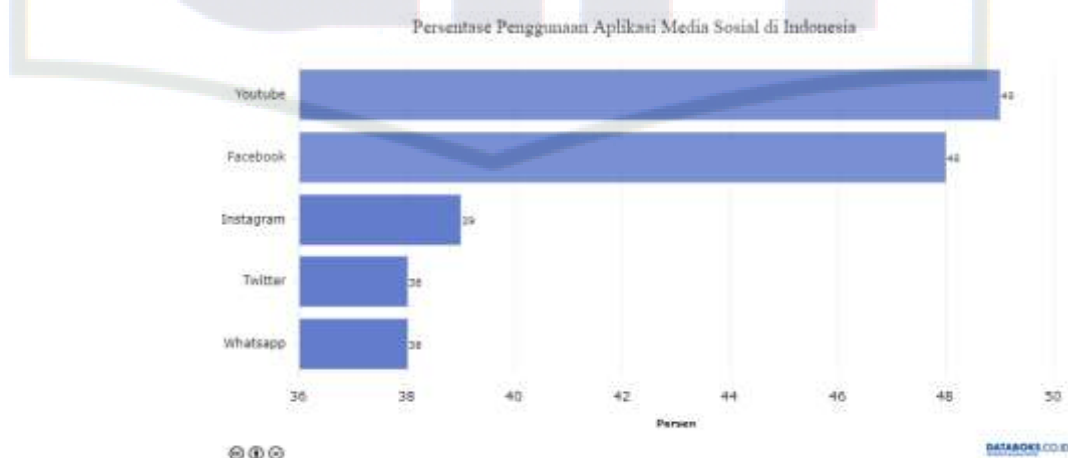
Adapun korelasi antara tujuan pemerintah dalam kebijakan ini dengan firman Allah SWT pada surat An-Nisa ayat 58 adalah sebagai berikut:

“Sesungguhnya Allah menyuruhmu menyampaikan amanah kepada yang berhak menerimanya. Dan apabila kamu menetapkan hukum di antara manusia, hendaknya kamu menetapkannya dengan adil.

Sesungguhnya Allah sebaik-baik yang memberi pengajaran kepadamu. Sungguh, Allah Maha Mendengar lagi Maha Melihat.”

Perkembangan teknologi digital yang semakin pesat, telah melahirkan berbagai media sosial sebagai sarana yang digunakan masyarakat untuk berkomunikasi dengan keluarga, teman sekolah atau rekan kerja bahkan untuk menambah teman dari latar belakang yang beraneka ragam, ataupun mengomentari berupa kritik dan saran terhadap kebijakan pemerintah seperti yang sedang dalam penelitian saat ini, khususnya tentang aturan zonasi sekolah melalui berbagai sosial media. Terbukti, populasi penduduk Indonesia saat ini mencapai 262 juta orang. Lebih dari 50% atau sekitar 146 juta orang telah terhubung jaringan internet sepanjang 2017 menurut laporan terbaru dari Asosiasi Penyelenggara Jasa Internet Indonesia (APJII).

Berdasarkan riset *We Are Social* dan *Hootsuite* pada Januari 2017, mengukuhkan Youtube sebagai aplikasi media sosial yang paling sering digunakan di Indonesia. Youtube berhasil mengungguli Facebook yang saat ini berada di posisi kedua. Tahun ini, setidaknya ada 100 juta orang mengakses Youtube dan diprediksi pengguna Youtube akan meningkat hingga tujuh kali lipat pada 2020. Maka dari itu, penulis memutuskan untuk mengambil data komentar dari Youtube dan Facebook. Presentase penggunaan sosial media di Indoneisa dapat dilihat pada gambar 1.1.



Gambar 1.1 Presentase Penggunaan Aplikasi Media Sosial di Indonesia

(Sumber : databoks.co.id)

Pada penelitian sebelumnya yang berkaitan dengan analisis sentiment yaitu penelitian dari (Nugroho, 2016) dengan topik “Analisis Sentimen Data Twitter Menggunakan *K-Means Clustering*”. Dipenelitian ini, peneliti mengambil data sebanyak 200 *tweets* dari twitter secara *random* dengan *keyword* hastag “#” antara lain: #cinta, #sedih, #senang, #marah, dan #takut. Peneliti melakukan *pre-processing* data terlebih dahulu seperti *tokenization*, *stopword removal*, dan *stemming* (menggunakan algoritma Nazief dan Adriani). Kemudian melakukan pembobotan kata menggunakan algoritma *TF-IDF*. Tahap terakhir untuk pengambilan sentimennya hanya menggunakan algoritma *k-means clustering* dengan $k=5$ karena menganalisis 5 emosi dan menghasilkan nilai *accuracy* sebesar 76,3%.

Penelitian sejenis selanjutnya yang berkaitan dengan analisis sentiment dan juga menjadi acuan terhadap peneliti yang sekarang yaitu penelitian dari (Rosandhy, 2017) dengan topik “Sistem Analisis Sentimen pada Komentar Evaluasi Dosen di SION STIKOM Bali”. Pada penelitian ini, peneliti mengambil data komentar sebagai evaluasi dosen di SION STIKOM Bali sebanyak 400 data. Peneliti melakukan *pre-processing* data terlebih dahulu seperti *case folding*, *tokenization*, *stopword removal*, dan *stemming* (menggunakan algoritma Nazief dan Adriani). Kemudian melakukan pembobotan kata menggunakan algoritma *TF-IDF*. Tahap terakhir untuk pengambilan sentimennya hanya menggunakan algoritma *k-means clustering* dengan $k=2$ karena output yang dicari hanya positif dan negatif dan menghasilkan nilai *accuracy* sebesar 85%.

Pada penelitian sejenis selanjutnya yaitu penelitian dari (Antinasari, Perdana, & Fauzi, 2017) dengan topik “Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku”. Dipenelitian ini, peneliti mengambil data dari hasil pencarian di twitter dengan bahasa Indonesia. Pada penelitian ini, peneliti menambahkan algoritma sebagai normalisasi

kata tidak baku menjadi kata baku dengan menggunakan algoritma *Levenshtein Distance*. Pada tahap terakhir untuk pengambilan sentimennya menggunakan algoritma *Naïve Bayes Classifier* menghasilkan nilai *accuracy* yang meningkat yaitu sebesar 98,33%.

Disebutkan juga pada penelitian (Antinasari et al., 2017), *levenshtein distance* atau yang biasa disebut dengan *edit distance* adalah suatu metode yang dapat digunakan untuk mengatasi terjadinya kesalahan ejaan. Kesalahan ejaan terjadi apabila kata yang diketik oleh pengguna tidak terdapat pada daftar kamus Bahasa Indonesia. Fungsi metode Levenshtein Distance yaitu untuk menghitung jarak kedekatan dari dua buah string melalui penambahan karakter, pengubahan karakter, dan penghapusan karakter hingga kedua string tersebut cocok.

Dari penelitian sebelumnya, penulis memutuskan untuk mengambil beberapa hal dari penelitian sebelumnya untuk diterapkan pada penelitian sekarang sehingga menjadikan penelitian penulis berbeda dan memiliki keunikan tersendiri dari penelitian lainnya. Penulis akan mengkombinasikan antara algoritma *K-Means* sebagai penentu sentimennya dan algoritma *Levenshtein Distance* sebagai algoritma normalisasi kata pada tahap *per-processing*-nya.

Alasan penulis mengkombinasikan 2 algoritma ini yaitu pada algoritma *K-Means* merupakan algoritma *clustering* dengan kompleksitas yang tidak sulit, implementasi juga mudah, dan relatif lebih efisien dalam proses implementasinya. Berbeda dengan algoritma *K-Medoid* yang juga merupakan algoritma *clustering*, namun *K-Medoid* ini memiliki tingkat kompleksitas yang lebih rumit, dan kurang efisien pada implementasinya. Pada algoritma *Levenshtein Distance* merupakan algoritma yang sering digunakan dalam pengejaan kata yang salah. Terbukti pada penelitian (Antinasari et al., 2017), algoritma *Levenshtein Distance* ini dapat meningkatkan nilai akurasi pada penelitiannya.

Berdasarkan permasalahan yang ada dilatar belakang di atas, penulis bermaksud untuk melakukan penelitian dengan topik “Analisis Sentimen

Komentar Masyarakat Terhadap Kebijakan Pemerintah Tentang Sistem Zonasi Sekolah Menggunakan Algoritma *K-Means Clustering* dan Algoritma *Levensthein*".

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka penulis merumuskan masalah bagaimana analisis sentimen terhadap sistem zonasi sekolah untuk melihat hasil tingkat akurasi penggunaan algoritma *K-Means Clustering* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein* dalam analisis sentimen terhadap kebijakan pemerintah tentang Zonasi Sekolah?

1.3 Batasan Masalah

Batasan-batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Komentar publik yang ingin dilihat analisis sentimennya yaitu tentang Zonasi Sekolah diambil dari komentar *Facebook Page* Kemendikbud RI dan *Channel YouTube* CNN Indonesia.
2. Teknik yang digunakan oleh penulis yaitu menggunakan teknik kombinasi antara algoritma *clustering* dan algoritma normalisasi kata. Sentimen publik yang dicari terhadap kebijakan pemerintah mengenai Zonasi Sekolah yaitu positif dan negatif menggunakan Metode *Clustering K-Means* dan dibantu oleh algoritma *Levensthein* sebagai normalisasi kata.
3. Dalam penggunaan algoritma *K-Means*, penulis menggunakan $k=2$.
4. Dalam pengumpulan data atau *Crawling data* komentar kebijakan pemerintah mengenai Zonasi Sekolah dari *Facebook Page* Kemendikbud RI secara manual dan *Channel YouTube* CNN Indonesia dilakukan dengan menggunakan Bahasa Pemrograman R.
5. Data yang didapatkan dan digunakan yaitu sebanyak 200 komentar.
6. Data untuk data latih sebanyak 75% dari jumlah komentar yang didapat yaitu 150 komentar.

7. Data untuk data uji sebanyak 25% dari jumlah komentar yang didapat yaitu 50 komentar.
8. Dalam *stemming* kata atau penguraian kata menjadi kata baku, penulis menggunakan algoritma Nazief dan Adriani.
9. Dalam fitur pembobotan kata, penulis menggunakan algoritma TF-IDF.
10. Dalam perhitungan tingkat akurasi, penulis menggunakan *Confusion Matrix*.
11. Sistem Analisis Sentimen komentar kebijakan pemerintah mengenai Zonasi Sekolah dari *Facebook Page* Kemendikbud RI dan *Channel YouTube* CNN Indonesia akan dilakukan melalui aplikasi berbasis *website* dengan menggunakan Bahasa Pemrograman PHP dan MySQL sebagai database.
12. Metode implementasi yang digunakan yaitu metode simulasi.

1.4 Tujuan Masalah

Adapun tujuan yang akan dicapai dalam penelitian ini adalah menganalisa tingkat akurasi penggunaan algoritma *K-Means Clustering* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein* dalam analisis sentimen terhadap kebijakan pemerintah tentang Zonasi Sekolah.

1.5 Manfaat Penelitian

Adapun manfaat yang didapat dari hasil penelitian ini adalah :

1.5.1 Bagi Penulis

- a) Dapat menerapkan ilmu yang didapat selama masa perkuliahan khususnya dalam pengimplementasian algoritma pada suatu sistem.
- b) Sebagai salah satu persyaratan untuk meraih gelar sarjana S1 Teknik Informatika UIN Syarif Hidayatullah Jakarta.

1.5.2 Bagi Universitas

- a) Mengukur tingkat kemampuan mahasiswa akan ilmu yang telah didapatkan selama masa perkuliahan.
- b) Menjadi bahan tolak ukur evaluasi kedepannya.

1.5.3 Bagi Pemerintah

- a) Sebagai bahan pertimbangan untuk kedepannya dari hasil penelitian yang telah dilakukan.
- b) Membantu mengevaluasi kebijakan yang telah ditetapkan demi kemaslahatan bersama.

1.6 Metodologi Penelitian

Metode yang digunakan penulis dalam menyusun penelitian ini dibagi menjadi 2 metode, yaitu metode pengumpulan data dan metode pengembangan sistem. Berikut penjelasannya:

1.6.1 Metode Pengumpulan Data

Dalam pengumpulan data yang digunakan oleh penulis dalam melakukan penelitian ini menggunakan 3 metode pengumpulan data, yaitu:

- a) Studi lapangan : observasi dan wawancara
- b) Studi pustaka : url, buku, studi literatur, dan jurnal

1.6.2 Metode Implementasi

Pada penelitian ini, penulis melakukan implementasi penelitian dengan menggunakan metode Simulasi. Adapun langkah-langkah yang dilakukan di dalam metode simulasi ini, yaitu:

1. Formulasi Masalah (*Problem Formulation*)
2. Model Pengkonsepkan (*Conceptual Model*)
3. Data Masukan/Keluaran (*Input/Output Data*)
4. Pemodelan (*Modelling*)
5. Simulasi (*Simulation*)
6. Verifikasi dan Validasi (*Verification and Validation*)
7. Eksperimentasi (*Experimentation*)

8. Analisis Keluaran (*Output Analysis*)

1.7 Sistematika Penulisan

Untuk memudahkan dalam penulisan laporan tugas akhir ini, penulis menyusunnya ke dalam beberapa bagian. Setiap babnya terdiri dari beberapa sub bab tersendiri. Dimana bab tersebut secara keseluruhan saling berkaitan satu sama lain. Berikut penjelasan singkat dari masing-masing bab:

BAB 1 PENDAHULUAN

Pada bab ini peneliti menjelaskan terkait latar belakang dari dari sebuah permasalahan yang diangkat, tujuan penelitian, manfaat penelitian, rumusan masalah, batasan masalah, metodologi penelitian, dan sistematika penulisan pada tugas skripsi ini.

BAB 2 LANDASAN TEORI

Pada bab ini peneliti menjelaskan tentang materi-materi apa saja yang dipakai untuk dijadikan dasar penelitian yang sedang dilakukan.

BAB 3 METODE PENELITIAN

Pada bab ini peneliti menjelaskan tentang metode penelitian apa yang dipakai untuk mendapatkan data dan metode untuk pengembangan sistem yang telah dibuat serta kerangka berpikir pembuatan tugas akhir ini.

BAB 4 IMPLEMENTASI, SIMULASI, DAN EKSPERIMEN

Pada bab ini menjelaskan tentang implementasi dari metode yang telah digunakan untuk perancangan membangun sebuah sistem dan tahapan proses menganalisa simulasi.

BAB 5 HASIL DAN PEMBAHASAN

Pada bab ini peneliti membahas tentang hasil yang telah didapat dari proses simulasi yang telah dilakukan pada bab sebelumnya.

BAB 6 PENUTUP

Pada bab ini peneliti menjelaskan tentang kesimpulan dari hasil yang telah didapat dan menjawab semua pokok permasalahan yang dirancang serta saran-saran yang digunakan untuk penelitian lebih lanjut.



BAB 2

LANDASAN TEORI

2.1 Analisis Sentimen

Analisis sentimen adalah bidang studi yang menganalisis pendapat, sentiment, evaluasi, penilaian, sikap, dan emosi seseorang terhadap sebuah produk, organisasi, individu, masalah, peristiwa atau topik. (Liu, 2012).

2.1.1 Level Analisis Sentimen

Analisis sentimen terdiri dari tiga level analisis, yaitu (Nugroho, 2016):

1. Level Dokumen

Level dokumen menganalisis satu dokumen penuh dan mengklasifikasikan dokumen tersebut memiliki sentiment positif atau negatif. Level analisis ini berasumsi bahwa keseluruhan dokumen hanya berisi opini tentang satu entitas saja. Level analisis ini tidak cocok diterapkan pada dokumen yang membandingkan lebih dari satu entitas.

2. Level Kalimat

Level kalimat menganalisis satu kalimat dan menentukan tiap kalimat bernilai sentimen positif, negatif, atau netral. Sentimen netral berarti kalimat tersebut bukan opini.

3. Level Entitas dan Aspek

Level aspek tidak melakukan analisis pada konstruksi Bahasa (dokumen, paragraf, kalimat, klausa, atau frase) melainkan langsung pada opini itu sendiri. Hal ini didasari bahwa opini terdiri dari sentimen (positif atau negatif) dan target dari opini tersebut. Tujuan level analisis ini adalah untuk menemukan sentimen entitas pada tiap aspek yang dibahas.

2.2 Sistem Zonasi Sekolah

Sistem zonasi, menurut Menteri Pendidikan dan Kebudayaan, Muhadjir Effendy, merupakan bentuk penyesuaian kebijakan dari sistem sebelumnya, yakni sistem rayonisasi. Rayonisasi lebih memperhatikan pada capaian siswa di bidang akademik, sementara sistem zonasi lebih menekankan pada jarak/radius antara rumah siswa dan sekolah. Dengan demikian, yang lebih berhak mendapatkan layanan pendidikan adalah calon siswa yang rumahnya paling dekat dengan sekolah (Kemendikbud, 2018).

2.3 Clustering

Clustering adalah proses pengelompokan benda serupa ke dalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah data set kedalam subset, sehingga data dalam setiap subset memiliki arti yang bermanfaat. Sebuah cluster terdiri dari kumpulan benda-benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada cluster lainnya.

Clustering juga bisa dikatakan suatu proses dimana mengelompokkan dan membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokkan pada *cluster* yang sama dan memisahkan diri dengan membentuk pola yang berbeda di cluster yang berbeda (Merliana, Ernawati, & Santoso, 2015).

2.4 Algoritma

Menurut Menurut (Sjukani, 2014), algoritma pada dasarnya, adalah alur pikiran dalam menyelesaikan suatu pekerjaan, yang dituangkan dalam bentuk tertulis yang dapat dimengerti oleh orang lain. yang ditekankan disini adalah alur pikiran. Alur pikiran seseorang dapat berbeda dengan alur pikiran orang lain untuk menyelesaikan suatu pekerjaan yang sama dengan hasil yang sama. Dalam bentuk tertulis, maksudnya dapat berupa narasi dalam bentuk kalimat, dapat juga berbentuk gambar atau bagan atau dalam bentuk tabel.

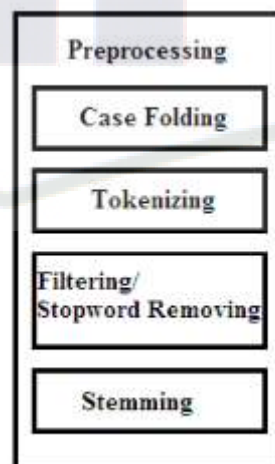
2.5 Text Mining

Text mining didefinisikan sebagai proses pengetahuan intensif yang melibatkan interaksi pengguna dengan sekumpulan dokumen dari waktu ke waktu menggunakan berbagai macam analisis. Sejalan dengan dengan *data mining*, *text mining* berusaha mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi *pattern* (Nugroho, 2016).

2.6 Pre-processing

Menurut (Jumeilah, 2017), dokumen pada umumnya mempunyai struktur yang sembarangan atau tidak terstruktur. Oleh karena itu, diperlukan suatu proses yang dapat mengubah bentuk data yang sebelumnya tidak terstruktur ke dalam bentuk data yang terstruktur. Proses pengubahan ini dikenal dengan istilah text preprocessing.

Proses preprocessing dilakukan agar data yang digunakan bersih dari noise, memiliki dimensi yang lebih kecil, serta lebih terstruktur, sehingga dapat diolah lebih lanjut. Tahap preprocessing memiliki beberapa proses, yaitu case folding, stopwords removing, tokenizing, dan stemming. Untuk lebih jelasnya dapat dilihat pada gambar 2.3.



Gambar 2.1 Tahap *Pre-processing*

2.6.1 *Case Folding*

Case Folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf a sampai z yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (Salim & Anistyasari, 2017).

2.6.2 *Tokenizing*

Adapun pengertian *tokenization* menurut (Salim & Anistyasari, 2017) adalah tahap pemotongan *string* masukan berdasarkan kata-kata yang menyusunnya atau dengan kata lain pemecahan kalimat menjadi kata. Strategi umum yang digunakan pada tahap *tokenizing* adalah memotong kata pada *white space* atau spasi dan membuang karakter tanda baca. Tahap *tokenizing* membagi urutan karakter menjadi kalimat dan kalimat menjadi *token*.

2.6.3 *Filtering/Stopword*

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Terdapat beberapa algoritma dalam filtering yaitu Stop-list dan word-list. Algoritma stop-word merupakan algoritma yang digunakan untuk mengeliminasi kata-kata yang tidak deskriptif. Algoritma word-list adalah algoritma yang digunakan menyimpan kata-kata memiliki nilai deskriptif (Salim & Anistyasari, 2017).

2.6.4 *Stemming*

Stemming merupakan proses untuk mendapatkan *root/stem* atau kata dasar dari suatu kata dalam kalimat dengan cara memisahkan masing-masing kata dari kata dasar dan imbuhanannya baik awalan (prefiks) maupun akhiran (sufiks). Sebagai contoh, kata bersama, kebersamaan, menyamai, akan di stem ke *root word* nya yaitu “sama”. (Wahyudi, Susyanto, & Nugroho, 2017)

Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma *stemming* untuk bahasa lainnya. Sebagai contoh Bahasa

Inggris memiliki morfologi yang berbeda dengan Bahasa Indonesia sehingga algoritma *stemming* untuk kedua bahasa tersebut juga berbeda. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks. Sedangkan pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan root word dari sebuah kata.

Efektifitas algoritma stemming dapat dipengaruhi oleh beberapa faktor (Wahyudi et al., 2017):

- a. Kesalahan dalam proses pemenggalan imbuhan dari kata dasarnya. Kesalahan ini dapat berupa: *Overstemming*, *Understemming*, *Unchange* dan *Spelling exception*.
- b. Kekurangan dalam perumusan aturan penambahan imbuhan pada kata dasar.
- c. Jumlah total aturan imbuhan yang didapat berhubungan dengan efektifitas proses temu kembali.

2.7 Algoritma Nazief dan Adriani

2.7.1 Tahapan Algoritma Nazief dan Adriani

Algoritma yang dibuat oleh Bobby Nazief dan Mirna Adriani memiliki tahapan sebagai berikut (Prasidhatama & Suryaningrum, 2018):

1. Cari kata yang akan distem dalam kamus. Jika ditemukan maka diasumsikan bahwa kata tersebut adalah kata dasar. Maka algoritma berhenti.
2. Infleksi akhiran (*Inflectional suffixes*) (“-lah”, “-kah”, “-ku”, “-mu”, atau “-nya”) dibuang. Jika berupa partikel (“lah”, “-kah”, “-tah” atau “-pun”) maka langkah ini diulangi lagi untuk menghapus kata ganti posesif (“-ku”, “-mu”, atau “-nya”), jika ada.

3. Hapus penurunan akhiran (*Derivational Suffix*) (“-i”, “-an” atau “kan”). Jika kata ditemukan di kamus, maka algoritma berhenti. Jika tidak maka ke langkah 3a.
 - a. Jika “-an” telah dihapus dan huruf terakhir dari kata tersebut adalah “-k”, maka “-k” juga ikut dihapus. Jika kata tersebut ditemukan dalam kamus maka algoritma berhenti. Jika tidak ditemukan maka lakukan langkah 3b.
 - b. Akhiran yang dihapus (“-i”, “-an” atau “kan”) dikembalikan, lanjut ke langkah 4.
4. Hapus penurunan awalan (*Derivational Prefix*) (“-di”, “ke-”, “se-”, “te-”, “be-”, “me-”, atau “pe-”). Jika pada langkah 3 ada akhiran yang dihapus maka pergi ke langkah 4a, jika tidak pergi ke langkah 4b.
 - a. Periksa tabel kombinasi awalan akhiran yang tidak diizinkan. Jika ditemukan maka algoritma berhenti, jika tidak pergi ke langkah 4b.
 - b. Pada langkah ini dilakukan perulangan sebanyak tiga kali. Tentukan tipe awalan kemudian hapus awalan. Jika kata dasar belum juga ditemukan lakukan langkah 5, jika sudah maka algoritma berhenti. Catatan: jika awalan kedua sama dengan awalan pertama algoritma berhenti.
5. Melakukan *Recoding*.

Jika semua langkah telah selesai tetapi tidak juga berhasil maka kata awal diasumsikan sebagai kata dasar lalu proses diakhiri.

2.7.2 Alasan Menggunakan Algoritma Nazief dan Adriani

Menurut penelitian (Simarangkir, 2017), dari hasil pengujian antara algoritma *stemming* Nazief & Adriani, Arifin & Setiono, Vega, dan Tala, tingkat akurasi tertinggi yaitu algoritma Nazief dan Adriani.

Tabel 2.1 Tingkat Akurasi Algoritma-Algoritma *Stemming*

Algoritma	Waktu Proses (detik)	Akurasi (%)
Nazief & Adriani	5,147	97,931
Arifin & Setiono	15,204	92,099
Vega	0,085	63,486
Tala	0,22	78,274

Dan dari penelitiannya itu menyimpulkan bahwa untuk algoritma yang menggunakan kamus, ditemukan algoritma terbaik dalam proses stemming yaitu algoritma Nazief dan Adriani. Hal ini dikarenakan pada algoritma Nazief dan Adriani terdapat penambahan aturan-aturan untuk reduplikasi, penambahan aturan untuk awalan dan akhiran dalam meningkatkan presisi dari setiap kata yang distemming.

2.8 Algoritma *K-Means Clustering*

2.8.1 Penjelasan Algoritma *K-Means Clustering*

K-Means merupakan salah satu metode data *clustering non-hierarchical* yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/kelompok. Metode ini mempartisi data ke dalam *cluster*/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam *cluster* lain (Widodo & Wahyuni, 2017).

2.8.2 Tahapan Algoritma *K-Means Clustering*

Pada penelitian (Rohmawati, Defiyanti, & Jajuli, 2015), algoritma *k-means clustering* memiliki beberapa tahapan seperti berikut:

1. Menentukan *k* sebagai jumlah *cluster* yang ingin dibentuk.
2. Membangkitkan nilai *random* untuk pusat *cluster* awal (*centroid*) sebanyak *k*.

3. Menghitung jarak setiap data *input* terhadap masing-masing *centroid* menggunakan rumus jarak *Euclidean* (*Euclidean Distance*) hingga ditemukan jarak yang paling dekat dari setiap data dengan *centroid*. Berikut adalah persamaan *Euclidian Distance*:

$$d(x_i, \mu_j) = \sqrt{\sum (x_i - \mu_j)^2} \quad (2.1)$$

Keterangan:

x_i : data kriteria

μ_j : *centroid* pada *cluster* ke- j

4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
5. Memperbaharui nilai *centroid*. Nilai *centroid* baru diperoleh dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus:

$$\mu_j(t+1) = \frac{1}{N_{sj}} \sum_{j \in S_j} x_j \quad (2.2)$$

Keterangan:

$\mu_j(t+1)$: *centroid* baru pada iterasi ke $(t+1)$

N_{sj} : banyak data pada *cluster* S_j .

6. Melakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap *cluster* tidak ada yang berubah.

Jika langkah 6 telah terpenuhi, maka nilai pusat *cluster* (μ_j) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

2.9 Algoritma *Levenshtein Distance*

2.8.1 Penjelasan Algoritma *Levenshtein Distance*

Menurut (Pratama & Pamungkas, 2016) *Levenshtein distance* adalah sebuah matriks string yang digunakan untuk mengukur perbedaan atau jarak (*distance*) antara dua string. Nilai *distance* antara dua string ini ditentukan oleh jumlah minimum dari

operasi-operasi perubahan yang diperlukan untuk melakukan transformasi dari suatu string menjadi string lainnya. Operasi-operasi tersebut adalah penyisipan (*insertion*), penghapusan (*deletion*), atau penukaran (*substitution*). Levenshtein distance merupakan salah satu algoritma yang dapat digunakan dalam mendeteksi kemiripan antara dua string yang berpotensi melakukan tindak plagiarism.

Pada algoritma *levenshtein distance* terdapat 3 macam operasi utama yang dilakukan, yaitu (Antinasari et al., 2017):

1. Operasi Penambahan Karakter

Operasi penambahan karakter yaitu operasi yang digunakan untuk menambahkan karakter ke dalam *string*. Contoh pada penulisan *string* “kern” maka diubah menjadi *string* “keren” dengan menambahkan karakter ‘e’.

2. Operasi Pengubahan Karakter

Operasi pengubahan karakter yaitu operasi yang digunakan untuk mengubah karakter dengan cara menukar sebuah karakter dengan karakter lain. Contoh pada penulisan *string* “hidsp” diubah menjadi *string* “hidup” dengan mengubah karakter ‘S’ menjadi karakter ‘U’.

3. Operasi Penghapusan Karakter

Operasi penghapusan karakter yaitu operasi yang digunakan untuk menghapus suatu karakter pada *string*. Contoh pada penulisan *string* “hebatt” diubah menjadi *string* “hebat” dengan menghilangkan salah satu karakter ‘T’.

2.8.2 Tahapan Algoritma *Levenshtein Distance*

Menurut (Junedy, 2014) dalam (Pratama B. P., 2016) algoritma Levenshtein distance berjalan mulai dari pojok kiri atas sebuah array dua dimensi (matriks) yang telah diisi sejumlah karakter string awal dan string target. Entri-entri pada matriks

tersebut merepresentasikan nilai terkecil dari transformasi string awal menjadi string target. Entri yang terdapat pada ujung kanan bawah matriks adalah nilai distance yang menggambarkan jumlah perbedaan dua string. Berikut ini adalah langkah-langkah algoritma *levenshtein distance* dalam mendapatkan nilai *distance*:

Misalkan:

S = String Awal

T = String Target

1. Inisialisai (nilai awal)
 - a. Hitung panjang S dan T, misalkan m dan n.
 - b. Buat matriks berukuran 0...m baris dan 0...n kolom.
 - c. Inisialisasi baris pertama dengan 0...n.
 - d. Inisialisasi kolom pertama dengan 0...m.
2. Proses perhitungan matriks
 - a. Periksa $S[i]$ untuk $1 < i < n$
 - b. Periksa $T[j]$ untuk $1 < j < m$
 - c. Jika $S[i] = T[j]$, maka entrinya adalah nilai yang terletak pada tepat didiagonal atas sebelah kiri, yaitu $d[i,j] = d[i-1,j-1]$ (2.3)
 - d. Jika $S[i] \neq T[j]$, maka entrinya adalah $d[i,j]$ minimum dari:
 - Nilai yang terletak tepat diatasnya, ditambah satu, yaitu $d[i,j-1]+1$. (2.4)
 - Nilai yang terletak tepat dikirinya, ditambah satu, yaitu $d[i-1,j]+1$. (2.5)
 - terletak pada tepat didiagonal atas sebelah kirinya, ditambah satu, yaitu $d[i-1,j-1]+1$. (2.6)
3. Hasil entri matriks pada baris ke-i dan kolom ke j, yaitu $d[i,j]$.
4. Langkah 2 diulang hingga entri $d[m,n]$ ditemukan.

Contoh tabel perhitungan:

Table 2.1 Perhitungan Matriks Algoritma *Levenshtein Distance*

		F	I	L	M
	0	1	2	3	4
F	1	0	1	2	3
L	2	1	1	1	2
M	3	2	2	2	1

2.10 Algoritma TF-IDF

Setiap *term* yang telah di-*index* diberikan bobot sesuai dengan struktur pembobotan yang dipilih, apakah pembobotan lokal, global atau kombinasi keduanya. Jika menggunakan pembobotan lokal maka, pembobotan *term* diekspresikan sebagai *tf* (*term frequency*). Namun, jika pembobotan global yang digunakan maka, pembobotan *term* didapatkan melalui nilai *idf* (*inverse document frequency*). Beberapa aplikasi juga ada yang menerapkan pembobotan kombinasi keduanya yaitu, dengan mengalikan bobot lokal dan global (*tf . idf*). (Bintana & Agustian, 2012)

1. *Term Frequency*

Empat cara yang dapat digunakan untuk memperoleh nilai term frequency (tf), yaitu:

- a. *Raw term frequency*. Nilai tf sebuah term diperoleh berdasarkan jumlah kemunculan term tersebut dalam dokumen. Contoh kasus dimana term muncul sebanyak dua kali dalam suatu dokumen maka, nilai tf term tersebut adalah 2.
- b. *Logarithm term frequency*. Hal ini untuk menghindari dominasi dokumen yang mengandung sedikit term dalam query, namun mempunyai frekuensi yang tinggi. Cara ini menggunakan fungsi logaritmik matematika untuk memperoleh nilai tf.

$$tf = 1 + \log(tf)$$
- c. *Binary term frequency*. Hanya memperhatikan apakah suatu term ada atau tidak dalam dokumen. Jika ada, maka tf diberi nilai 1, jika tidak ada diberi nilai 0. Pada cara ini jumlah kemunculan term dalam dokumen tidak berpengaruh.

- d. *Augmented Term Frequency*. Nilai tf adalah jumlah kemunculan suatu term pada sebuah dokumen, sedangkan nilai $\max(tf)$ adalah jumlah kemunculan terbanyak sebuah term pada dokumen yang sama.

$$tf = 0.5 + 0.5 \times \frac{tf}{\max(tf)}$$

2. *Inverse Document Frequency*

Inverse document frequency (idf) digunakan untuk memberikan tekanan terhadap dominasi *term* yang sering muncul di berbagai dokumen. Hal ini diperlukan karena *term* yang banyak muncul di berbagai dokumen, dapat dianggap sebagai *term* umum (*common term*) sehingga tidak penting nilainya. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu *term* (*inverse document frequency*).

$$idf(t) = \log\left(\frac{n}{df(t)}\right) \quad (2.7)$$

Keterangan:

n = jumlah dokumen dalam *corpus*

$df(t)$ = *document frequency* / jumlah dokumen dalam *corpus* yang mengandung *term* t .

3. *Term Weighting TF-IDF*

Menurut (Okfalisa & Harahap, 2016), apaun algoritma yang digunakan untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci (*query*) yaitu:

$$W(d, t) = tf(d, t) \cdot idf(t) \quad (2.8)$$

Keterangan:

n = jumlah dokumen dalam *corpus*

$df(t)$ = *document frequency* / jumlah dokumen dalam *corpus* yang mengandung *term* t .

2.11 Confusion Matrix

Pada data mining untuk mengukur kinerja dari model yang dihasilkan salah satunya menggunakan *confusion matrix*. *Confusion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data minin. (Rosandy, 2016)

Tabel 2.2 Model *Confusion Matrix*
(Subari & Ferdinandus, 2015)

	Positive (+)	Negative (-)
Positive (+)	True positive (<i>tp</i>)	False positive (<i>fp</i>)
Negative (-)	False negative (<i>fn</i>)	True negative (<i>tn</i>)

Formula menghitung akurasi dapat juga dituliskan sebagai berikut:

$$\frac{\text{Jumlah data yang diprediksi benar}}{\text{Total data yang diprediksi}} = \frac{(tp + tn)}{(tp + fp + tn + fn)} \times 100 \quad (2.9)$$

2.12 PHP

PHP merupakan secara umum dikenal sebagai bahasa pemrograman *script script* yang membuat dokumen HTML secara *on the fly* yang dieksekusi di server web, dokumen HTML yang dihasilkan dari suatu aplikasi bukan dokumen HTML yang dibuat dengan menggunakan editor teks atau editor HTML. Dikenal juga sebagai bahasa pemrograman *server-side*. (Sidik, 2014)

2.13 MySQL

MySQL merupakan *software* database yang termasuk paling populer di lingkungan Linux, kepopuleran ini karena ditunjang karena performansi query dari databasenya yang saat itu bisa dikatakan paling cepat, dan jarang bermasalah. MySQL telah tersedia juga di lingkungan Windows. Database MySQL kini telah dimiliki oleh Oracle.

Pengembangan MySQL kemudian mengembangkan database yang murni *opensource* dan *freeware* dengan nama MariaDB (Sidik, 2014)

2.14 *R-Programming*

R bukan saja bahasa tetapi juga lingkungan (*environment*) untuk komputasi statistik dan grafik. R merupakan project GNU yang dikembangkan oleh Bell Laboratories (sebelumnya AT&T, sekarang Lucent Technologies). Team pengembang R adalah John Chamber dan teman-temannya.

R menyediakan berbagai macam *tool* statistik dari linier dan memodelan non linier, uji statistik klasik, analisis *time-series*, klasifikasi, *clustering*, dan lain-lain. R juga menyediakan *tool* teknik grafis yang bertujuan untuk menampilkan data yang telah diolah secara visual dalam bentuk grafik.

R merupakan project *open-source* yang memungkinkan banyak pihak untuk memberikan kontribusi dalam proses pengembangan. (Faisal, 2017)

2.15 Metode Pengumpulan Data

Hal pertama yang dilakukan dalam analisis sitem adalah melakukan pengumpulan data. Ada beberapa teknik pengumpulan data yang dilakukan (A.S & Shalahuddin, 2014). Adapun teknik yang digunakan pada penelitian ini yaitu:

2.15.1 Teknik Wawancara

Pengumpulan data dengan menggunakan wawancara mempunyai beberapa keuntungan sebagai berikut.

1. Lebih mudah dalam menggali bagian sistem mana yang dianggap baik dan bagian mana yang dianggap kurang baik
2. Jika ada bagian tertentu yang menurut anda perlu untuk digali lebih dalam, anda dapat langsung menanyakan kepada narasumber.

3. Dapat menggali kebutuhan *user* secara lebih bebas.
4. *User* dapat mengungkapkan kebutuhannya secara lebih bebas.

Selain mempunyai beberapa kelebihan tersebut, teknik wawancara juga mempunyai beberapa kelemahan. Berikut ini adalah beberapa kelemahan dari teknik wawancara.

1. Wawancara akan sulit dilakukan jika narasumber kurang dapat mengungkapkan kebutuhannya.
2. Pertanyaan dapat menjadi tidak terarah, terlalu fokus pada hal-hal tertentu dan mengabaikan bagian lainnya.

Berikut ini adalah beberapa panduan dalam melakukan kegiatan wawancara agar memperoleh data yang diharapkan.

1. Buatlah jadwal wawancara dengan narasumber dan beritahukan maksud dan tujuan wawancara.
2. Buatlah panduan wawancara yang akan anda jadikan arahan agar pertanyaan dapat fokus kepada hal-hal yang dibutuhkan.
3. Gunakan pertanyaan yang jelas dan mudah dipahami.
4. Cobalah untuk menggali mengenai kelebihan dan kekurangan sistem yang telah berjalan sebelumnya.
5. Anda boleh berimprovisasi dengan mencoba menggali bagian-bagian tertentu yang menurut anda penting.
6. Catat hasil wawancara tersebut.

2.15.2 Teknik Observasi

Pengumpulan data dengan menggunakan observasi mempunyai keuntungan yaitu.

1. Analisis dapat melihat langsung bagaimana sistem lama berjalan
2. Mampu menghasilkan gambaran lebih baik jika dibanding dengan teknik lainnya.

Sedangkan kelemahan dengan menggunakan teknik observasi adalah

1. Membutuhkan waktu cukup lama
2. Orang-orang yang sedang diamati biasanya perilakunya akan berbeda dengan perilaku sehari-hari (cenderung berusaha terlihat baik).
3. Dapat mengganggu pekerjaan orang-orang pada bagian yang sedang diamati.

Berikut adalah beberapa petunjuk untuk melakukan teknik observasi

1. Tentukan hal-hal apa saja yang akan diobservasi agar kegiatan observasi menghasilkan sesuai dengan yang diharapkan.
2. Mintalah izin kepada orang yang berwenang pada bagian yang akan diobservasi
3. Berusaha sesedikit mungkin agar tidak mengganggu pekerjaan orang lain.
4. Jika ada yang tidak mengerti, cobalah bertanya. Jangan membuat asumsi sendiri.

2.15.3 Teknik Kuisisioner

Pengumpulan data yang menggunakan kuisisioner mempunyai kelebihan yaitu

1. Hasilnya lebih objektif, karena kuisisioner dapat dilakukan kepada banyak orang sekaligus.
2. Waktunya lebih singkat

Sedangkan kelemahan pengumpulan data dengan menggunakan kuisisioner adalah sebagai berikut

1. Responden cenderung malas untuk mengisi kuisisioner

2. Sulit untuk membuat pertanyaan yang singkat, jelas, dan mudah dipahami

Berikut ini adalah beberapa cara yang dapat dilakukan untuk membuat teknik kuisioner menghasilkan data yang baik

1. Hindari pertanyaan isian, lebih baik pilihan ganda.
2. Buatlah pertanyaan yang tidak terlalu banyak
3. Buatlah pertanyaan yang singkat, padat dan jelas

2.16 Metode Simulasi

Metode simulasi merupakan metode untuk melakukan simulasi dan pemodelan yang diadaptasi dari penelitian yang dilakukan oleh (Madani, Kazmi, & Mahlknecht, 2010) dengan judul “Wireless Sensor Networks: Modelling and Simulation”.

Menurut (Madani et al., 2010), yang dikutip dari skripsi (Saputro, 2016) metode simulasi terdiri dari beberapa tahapan yang terdiri dari:

2.16.1 Formulasi Masalah (*Problem Formulation*)

Proses simulasi dimulai dengan masalah praktis yang memerlukan pemecahan atau pemahaman. Sebagai contoh sebuah perusahaan kargo ingin mencoba untuk mengembangkan strategi baru untuk pengiriman truk, contoh lain yaitu astronom mencoba memahami bagaimana sebuah nebula terbentuk. Pada tahap ini kita harus memahami perilaku dari sistem, mengatur operasi sistem sebagai objek untuk percobaan. Maka kita perlu menganalisa berbagai solusi dengan menyelidik hasil sebelumnya dengan masalah yang sama. Solusi yang paling diterima yang harus dipilih.

2.16.2 Model Pengkonsepan (*Conceptual Model*)

Langkah ini terdiri dari deskripsi tingkat tinggi dari struktur dan perilaku sebuah sistem dan mengidentifikasi semua benda

dengan atribut dan interface mereka. Kita juga harus menentukan variabel state-nya, bagaimana cara mereka berhubungan, dan mana yang penting untuk penelitian. Pada tahap ini dinyatakan aspek-aspek kunci dari requirement. Selama definisi model konseptual, kita perlu mengungkapkan fitur yang penting. Kita juga harus mendokumentasikan informasi non-fungsional, misalnya seperti perubahan pada masa yang akan datang, perilaku nonintuitive atau non-formal, dan hubungan dengan lingkungan.

2.16.3 Data Masukan Keluaran (*Input Output Data*)

Pada tahap ini kita mempelajari sistem untuk mendapatkan data input dan output. Untuk melakukannya kita harus mengumpulkan dan mengamati atribut yang telah ditentukan pada tahap sebelumnya. Ketika entitas sistem yang dipelajari, maka dicoba mengaitkannya dengan waktu. Isu penting lainnya pada tahap ini adalah pemilihan ukuran sampel yang valid secara statistik dan format data yang dapat diproses dengan komputer. Kita harus memutuskan atribut mana yang stokastik dan deterministik. Dalam beberapa kasus, tidak ada sumber data yang dapat dikumpulkan (misalnya pada sistem yang belum ada). Dalam kasus tersebut kita perlu mencoba untuk mendapatkan set data dari sistem yang ada (jika tersedia). Pilihan lain yaitu dengan menggunakan pendekatan stokastik untuk menyediakan data yang diperlukan melalui generasi nomor acak.

2.16.4 Pemodelan (*Modelling*)

Pada tahap pemodel, kita harus membangun representasi yang rinci dari sistem berdasarkan model konseptual dan input/output data yang dikumpulkan. Model ini dibangun dengan mendefinisikan objek, atribut, dan metode menggunakan paradigma yang dipilih. Pada tahap ini spesifikasi model dibuat, termasuk set persamaan yang mendefinisikan perilaku dan

struktural. Setelah menyelesaikan definisi ini, kita harus membangun struktur awal model (mungkin berkaitan sistem dan metrik kerja).

2.16.5 Simulasi (*Simulation*)

Pada tahap simulasi, kita harus memilih mekanisme untuk menerapkan model (dalam banyak kasus menggunakan komputer dan bahasa pemrograman dan alat-alat yang memadai), dan model simulasi yang dibangun. Selama langkah ini, mungkin perlu untuk mendefinisikan algoritma simulasi dan menerjemahkannya ke dalam program komputer.

2.16.6 Verifikasi dan Validasi (*Verification and Validation*)

Pada tahap sebelumnya, tiga model yang berbeda yang dibangun yaitu model konseptual (spesifikasi), sistem model (Desain), dan model simulasi (*executable program*). Kita perlu memverifikasi dan memvalidasi model ini. Verifikasi terkait dengan konsistensi internal antara tiga model. Validasi difokuskan pada korespondensi antara model dan realitas yaitu hasil simulasi yang konsisten dengan sistem yang dianalisis.

2.16.7 Experimentasi (*Experimentation*)

Kita harus menjalankan model simulasi, menyusul tujuan yang dinyatakan pada model konseptual. Selama fase ini kita harus mengevaluasi output dari simulator menggunakan korelasi statistik untuk menentukan tingkat presisi untuk metrik kerja. Fase ini dimulai dengan desain eksperimen, menggunakan teknik yang berbeda. Beberapa teknik ini meliputi analisis sensitivitas, optimasi, dan seleksi (dibandingkan dengan sistem alternatif).

2.16.8 Analisa Keluaran (*Output Analysis*)

Pada tahap analisa keluaran, keluaran simulasi dianalisis untuk memahami perilaku sistem. Keluaran ini digunakan untuk

mendapatkan tanggapan tentang perilaku sistem yang asli. Pada tahap ini, alat visualisasi dapat digunakan untuk membantu proses tersebut.

2.17 Studi Literatur Sejenis

Berikut adalah literatur sejenis:



Tabel 2.3 Studi Literatur Sejenis

Peneliti (Tahun Penelitian)	Gregorius Agung Purwanto Nugroho (2016)	Setyo Budi (2017)	Prananda Antinasari, dkk (2017)	Christian Rosandhy (2017)	Penelitian Penulis Sekarang
Judul Penelitian	Analisis Sentimen Data Twitter Menggunakan <i>K-Means Clustering</i>	<i>Text Mining</i> Untuk Analisis Sentimen Review Film Menggunakan <i>K-Means Clustering</i>	Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku	Sistem Analisis Sentimen Pada Komentar Evaluasi Dosen di SION STIKOM Bali Menggunakan Gabungan Metode <i>K-Means</i> dan <i>K-Nearest Neighbor</i>	Analisis Sentimen Tentang Kebijakan Pemerintah Sistem Zonasi Sekolah dari Kemendikbud dengan Menggunakan Algoritma <i>K-Means Clustering</i> dan Algoritma <i>Levensthein</i>
Data yang diambil sentimennya	Hasil pencarian twitter dengan menggunakan hashtag #cinta, #sedih, #senang, #marah, dan #takut	Dataset review film yang diunduh dari (http://www.cs.cornell.edu/People/pabo/movie-review-data/)	Hasil Pencarian twitter berbahasa Indonesia	Komentar Evaluasi Dosen di SION STIKOM Bali	Facebook Page Resmi Kemendikbud RI dan Youtube CNN Indonesia
Algoritma atau Metode yang digunakan	Algoritma <i>K-Means</i>	Algoritma <i>K-Means</i>	Algoritma Naïve Bayes	Algoritma <i>K-Means</i>	Algoritma <i>K-Means</i>
Nilai <i>k</i>	<i>k</i> = 5 (cinta, sedih, senang, marah, dan takut)	-	-	<i>k</i> = 2 (positif dan negatif)	<i>k</i> = 2 (positif dan negatif)
Tools / Bahasa pemrograman yang digunakan	Java	RapidMiner	-	PHP	<i>R Programming</i> dan PHP

Data yang digunakan	200 tweets random dari setiap hastag yang digunakan dengan total tweets yaitu 1000 tweets	2000 komentar Review Film bahasa Inggris	-	400 data komentar (200 data latih dan 200 data uji)	200 <i>data crawling</i> dari YouTube dan Facebook (150 data latih dan 50 data uji)
Stemming	Algoritma Nazief dan Adriani	Porter	-	Algoritma Nazief dan Adriani	Algoritma Nazief dan Adriani
Pembobotan	Algoritma (<i>TF-IDF</i>)	Algoritma (<i>TF-IDF</i>)	-	Algoritma (<i>TF-IDF</i>)	Algoritma (<i>TF-IDF</i>)
Normalisasi kata	-	-	Normalisasi Algoritma Levensthein	-	Normalisasi Algoritma Levensthein
Tahap Pengujian	Pengujian dilakukan 1 kali dengan 1000 tweets Hasilnya berupa sentimen Positif, Negatif, dan Netral	Pengujian dilakukan 1 kali Hasilnya berupa sentimen Positif dan Negatif	Pengujian dilakukan 3x yaitu dengan menggunakan <i>pre-proccesing</i> tanpa menggunakan perbaikan kata baku, menggunakan perbaikan kata baku tanpa menggunakan <i>pre-processing</i> , dan menggunakan keduanya Hasilnya berupa akurasi setiap tahap.	Hasilnya berupa akurasi dari penggunaan Algoritma <i>K-Means</i>	Pengujian dilakukan 2 kali dengan menggunakan Algoritma <i>K-Means</i> dan kombinasi Algoritma <i>K-Means</i> dan Algoritma <i>Levensthein</i>

Adapun perbedaan antara peneliti saat ini dengan peneliti sebelumnya yaitu :

1. Peneliti melakukan penelitian tentang analisis sentimen masyarakat terhadap kebijakan pemerintah sistem zonasi sekolah dari Kemendikbud.

2. Peneliti menggunakan kombinasi algoritma *k-means clustering* dan algoritma *levensthein distance* sebagai pembeda dari peneliti sebelumnya.
3. Peneliti melakukan 2x percobaan dalam tahap pengujian yaitu tahap pengujian menggunakan algoritma *k-means* saja dan tahap pengujian menggunakan kombinasi algoritma *k-means* dan algoritma *levensthein distance*



BAB 3

METODOLOGI PENELITIAN

3.1 Metode Pengumpulan Data

Pada penelitian ini penulis mengumpulkan data dan informasi sebagai penunjang kebutuhan dalam proses pembuatan sistem ini menggunakan studi lapangan dan studi pustaka.

3.2.1 Studi Lapangan

1. Teknik Observasi

Peneliti melakukan observasi atau pengamatan secara langsung dan mengambil data dari bulan Juni 2018 – Agustus 2018 melalui YouTube API dan Facebook secara manual yaitu komentar masyarakat mengenai kebijakan pemerintah tentang Zonasi Sekolah dari Facebook *Page* Kemendikbud RI dan *Channel* YouTube CNN Indonesia. Didapat data sebanyak 200 komentar. Setelah mendapatkan data, peneliti melakukan pelabelan sentimen secara manual menggunakan 150 data secara random sebagai data latih. Pada saat pelabelan peneliti dibantu oleh 4 orang mahasiswa lulusan S1 agar data tidak subjektif.

2. Teknik Wawancara

Peneliti melakukan wawancara pada tanggal 24 Oktober 2018 di Kemendikbud RI kepada 2 pihak dari Kemendikbud RI-nya langsung secara bersamaan yaitu ibu Any Sayekti selaku kepala bagian hukum, tatalaksana, dan kepegawaian dan ibu Sopha Julia selaku pegawai pada bagian tersebut.

3.2.2 Studi Pustaka

Proses pengumpulan data-data akurat dengan cara mengumpulkan literatur-literatur sejenis dari berbagai buku, jurnal

dan skripsi yang berkaitan dengan permasalahan yang dibahas oleh peneliti.

3.2 Metode Simulasi

Metode simulasi ini digunakan untuk melihat hasil sentimen masyarakat dari objek yang diteliti yaitu kebijakan pemerintah tentang zonasi sekolah dengan menggunakan algoritma *k-means clustering* dan algoritma *levenshtein distance* sebagai normalisasi kata menjadi kata baku. Pada metode simulasi ini meliputi beberapa langkah atau tahap yang akan dilakukan yaitu:

3.2.1 Formulasi Masalah (*Problem Formulation*)

Tahap formulasi masalah merupakan langkah awal dalam perancangan pada model metode simulasi. Formulasi masalah merupakan suatu kegiatan untuk melakukan identifikasi masalah berdasarkan hasil penelitian sebelumnya (pada tabel 2.). Dalam tahap formulasi masalah ini peneliti merumuskan sebuah masalah yaitu mengimplementasikan kombinasi algoritma *k-means clustering* dan algoritma *levenshtein distance* dalam proses penentuan sentimen masyarakat terhadap kebijakan pemerintah tentang zonasi sekolah. Pada penelitian-penelitian sebelumnya yaitu penelitian dari (Nugroho, 2016), (Budi, 2017), dan (Rosandhy, 2017), belum ada yang mengkombinasikan antara algoritma *k-means clustering* dan algoritma *levenshtein distance* sebagai algoritma pendukung untuk normalisasi kata baku. Adapun yang mengimplementasikan kombinasi dengan algoritma *levenshtein distance* yaitu penelitian dari (Antinasari et al., 2017), pada penelitiannya peneliti mengkombinasikan algoritma *naïve bayes* dan algoritma *levenshtein distance*.

3.2.2 Model Pengkonsep (Conceptual Model)

Pada tahapan ini peneliti membuat model konsep yang akan dilakukan yaitu membahas keseluruhan penelitian ini. Konsep

pertama membuat konsep pada proses *text mining* yang ingin digunakan. Kedua, dengan mengidentifikasi input pada penelitian ini, yaitu komentar-komentar masyarakat terkait kebijakan zonasi sekolah dari *channel* youtube CNN Indonesia dan facebook *page* Kemendikbud, kemudian komentar yang telah dikumpulkan kemudian diolah dan diproses dengan secara manual untuk pelabelan terhadap data latih. Ketiga, membuat konsep untuk tahap uji pada skenario 1 yaitu dengan melihat hasil sentimen dan tingkat akurasi menggunakan algoritma *K-Means* saja. Keempat, tahap uji pada skenario 2 yaitu melihat hasil sentimen dan tingkat akurasi menggunakan kombinasi algoritma *K-Means* dan dibantu algoritma *Levensthein Distance* sebagai normalisasi kata pada tahap *pre-processing*.

3.2.3 Data Masukan/Keluaran (*Input/Output Data*)

Data masukan seperti kamus kata dasar, kamus *stopword*, kamus untuk *levensthein*, dan data komentar yang didapat dari Youtube API dan Facebook API dijadikan input pada penelitian ini dalam aplikasi berbasis PHP. Data yang diambil sebanyak 200 komentar. Data terdiri dari 150 komentar dijadikan data latih dan 50 komentar dijadikan data uji. Data pada aplikasi ini diolah menggunakan algoritma *K-Means* dan algoritma *Levensthein Distance* untuk menghasilkan output berupa hasil sentimen akhir dan tingkat akurasi dari skenario 1 dan skenario 2.

3.2.4 Pemodelan (*Modelling*)

Pada tahapan ini peneliti menentukan model skenario yang akan digunakan. Pada tahap ini penulis melakukan pemodelan dalam membuat rancangan sistem yang akan dibuat secara manual. Pemodelan atau skenario yang dibuat yaitu skenario kombinasi antara algoritma *K-Means* dan algoritma *Levensthein Distance* dan

skenario tanpa menggunakan algoritma *Levensthein Distance* (hanya menggunakan algoritma *K-Means* saja).

3.2.5 Simulasi (Simulation)

Pada tahapan ini, sistem akan dijalankan untuk mensimulasikan kinerja masing-masing algoritma sesuai dengan konsep dan skenario yang telah ditentukan sebelumnya. Simulasi yang akan dilakukan adalah dengan melakukan input dataset latih dan uji, melakukan pelabelan terhadap data latih secara manual untuk dikelompokkan sentimennya, melakukan pelatihan terhadap data latih dan melakukan *clustering* data uji. Hasil simulasi berupa perbandingan akurasi dari algoritma yang dijadikan penelitian, kemudian akan dicatat dan kemudian dilakukan tahap verifikasi.

3.2.6 Verifikasi dan Validasi (Verification and Validation)

Pada tahapan ini peneliti melakukan verifikasi dan validasi dari tahap sebelumnya. Pada tahap verifikasi dilakukan untuk memastikan adanya kesalahan atau tidak yang terjadi dalam beberapa tahapan atau proses simulasi. Sedangkan tahapan validasi dilakukan untuk memastikan kesesuaian proses simulasi yang dibuat berdasarkan model pengkonsepkan dengan formulasi masalah yang dibuat.

Pada intinya, verifikasi dan validasi bertujuan untuk menyakinkan hasil dari aplikasi sentimen ini sesuai dengan yang dikonsepskan sebelumnya.

3.2.7 Eksperimentasi (Experimentation)

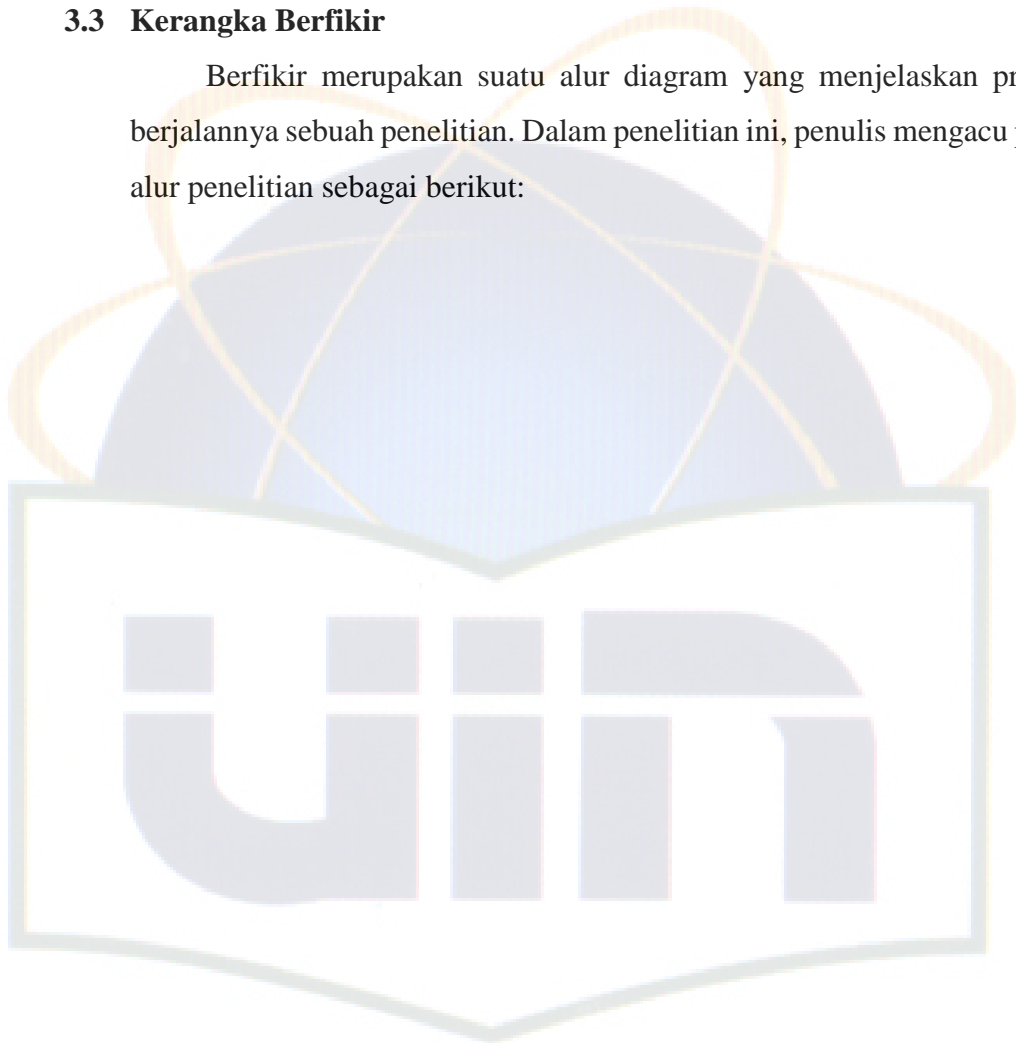
Pada tahapan ini peneliti melakukan eksperimen sesuai dengan model skenario yang dibuat pada saat tahap pemodelan. Eksperimen disini bertujuan untuk mengevaluasi hasil simulasi aplikasi.

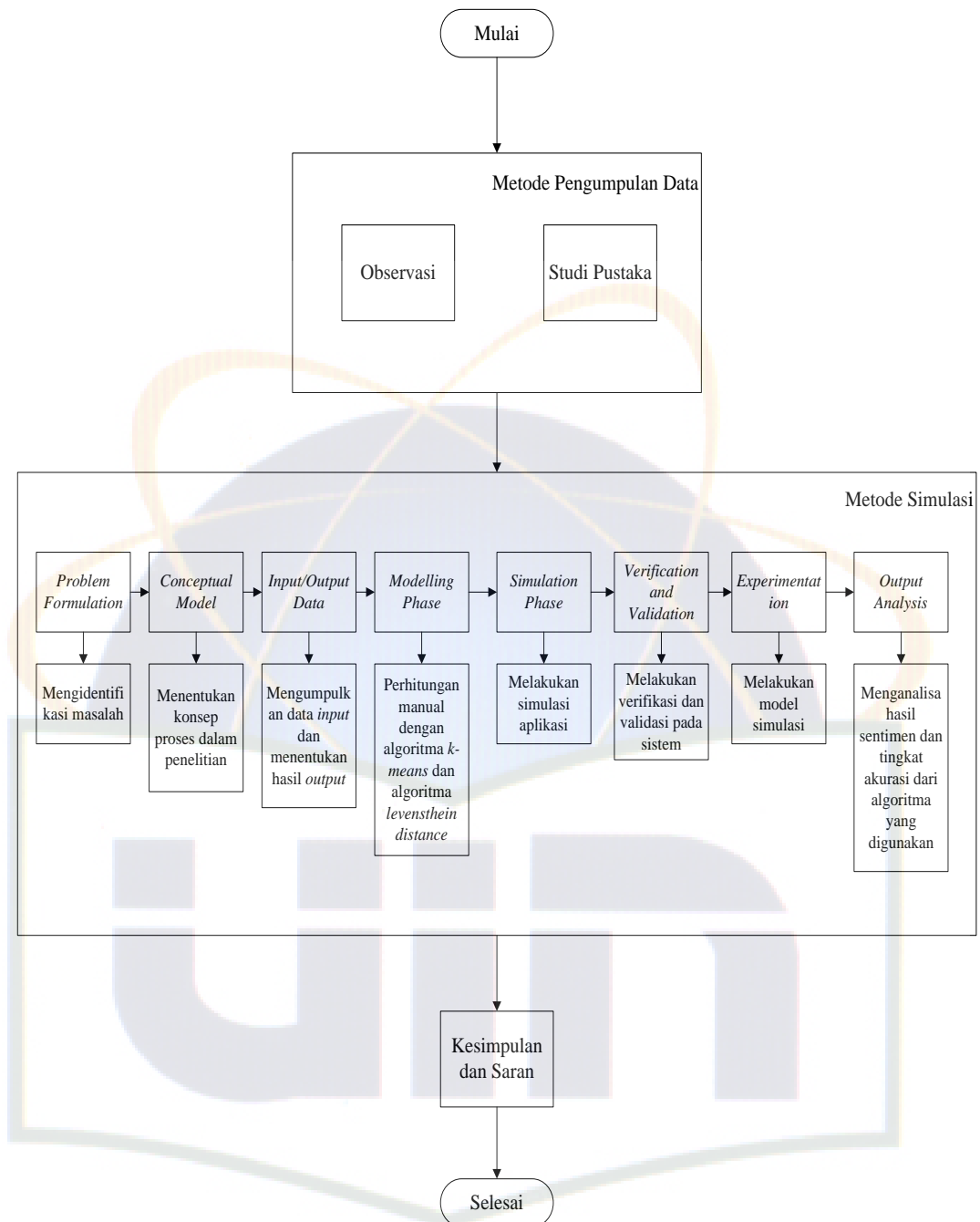
3.2.8 Analisis Keluaran (*Output Analysis*)

Tahapan analisis keluaran adalah tahapan simulasi yang paling terakhir. Peneliti melakukan analisa terhadap output-output berdasarkan skenario yang sudah dilakukan yaitu menghitung tingkat akurasi dari algoritma yang dijadikan penelitian.

3.3 Kerangka Berfikir

Berfikir merupakan suatu alur diagram yang menjelaskan proses berjalannya sebuah penelitian. Dalam penelitian ini, penulis mengacu pada alur penelitian sebagai berikut:





Gambar 3.1 Kerangka Berfikir

BAB 4

IMPLEMENTASI, SIMULASI, DAN EKSPERIMEN

4.1 Formulasi Masalah (*Problem Formulation*)

Pada tahap awal dimetode simulasi ini yaitu formulasi masalah, penulis melakukan identifikasi masalah berdasarkan hasil penelitian sebelumnya. Alhasil pada penelitian ini, penulis memformulasikan masalah penelitian pada kombinasi algoritma *k-means* dan algoritma normalisasi kata yaitu *levenshtein distance* untuk dilakukan kombinasi terhadap kedua algoritma tersebut. Dengan hasil akhirnya yaitu sentiment masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah dan tingkat akurasi dari algoritma tersebut.

Data yang digunakan pada penelitian ini adalah komentar berbahasa Indonesia tentang kebijakan pemerintah yaitu sistem zonasi sekolah yang diambil dari *channel* YouTube dan Facebook *page*.

4.2 Model Pengkonsepian (*Conceptual Model*)

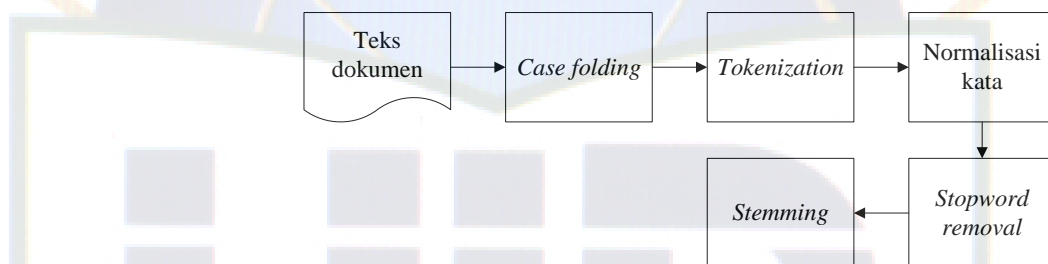
Pada tahap pemodelan konsep ini berkaitan dengan eksekusi dari sistem yang peneliti bangun, dari mulai masukan, proses, dan keluaran sistem tersebut. Berdasarkan tahap *conceptual model* pada sub bab 3.2.2, berikut ini merupakan alur keseluruhan dari sistem yang dibangun oleh peneliti.

4.2.1 *Conceptual Model* Pada *Text Mining*

Dalam penelitian saat ini, pemodelan pada *text mining* berkaitan dengan tahapan *pre-processing* teks. Tahapan *pre-processing* dilakukan dengan membuat fungsi sendiri dalam pengkodean menggunakan bahasa pemrograman PHP. *Pre-processing* ini dilakukan dengan tujuan:

1. Menghilangkan kata-kata yang mengganggu proses sentimen, seperti *url* atau *link*, hastag pada komentar, angka-angka, maupun tanda baca.
2. Menyeragamkan bentuk kata menjadi kata dasar sesuai dengan KBBI.
3. Mengurangi kata yang tidak digunakan atau tidak berpengaruh terhadap penentuan sentimen, seperti aku, kamu, dia, kita, dll.

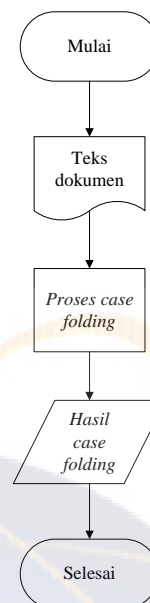
Adapun tambahan proses yang dilakukan pada penelitian ini sekaligus pembeda dari penelitian yang sebelum-sebelumnya di tahapan *pre-processing* ini yaitu proses normalisasi kata menggunakan algoritma *levensthein distance*, berikut tahap-tahap *pre-processing* pada penelitian saati ini meliputi proses-proses seperti yang digambarkan pada *flow* di bawah ini:



Gambar 4.1 Flowchart Tahapan Pre-Processing

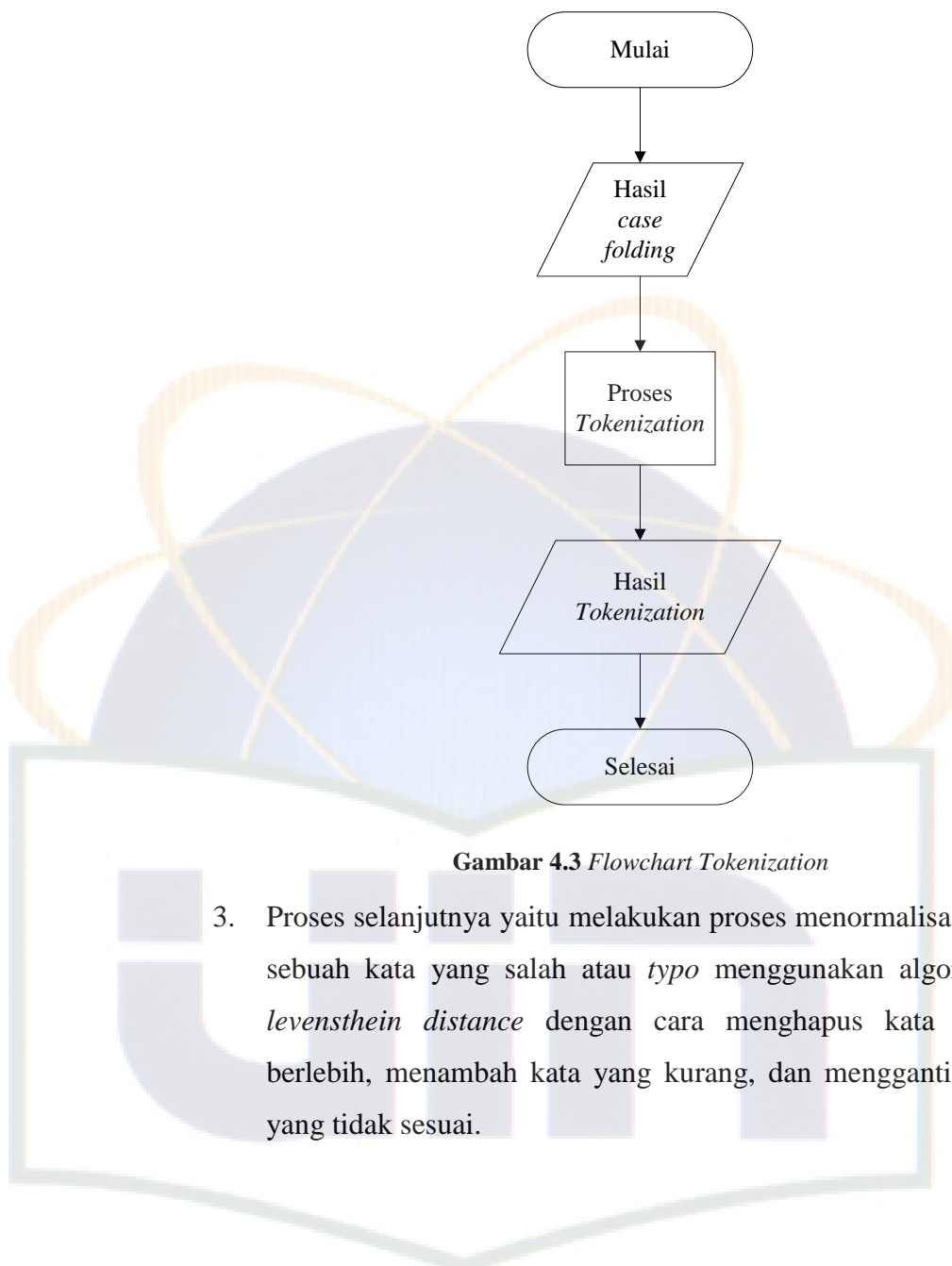
Berikut penjelasan dari tahapan *pre-processing* :

1. Tahap awal di *pre-processing* yaitu proses *case folding* atau merubah semua kata menjadi huruf kecil semua.



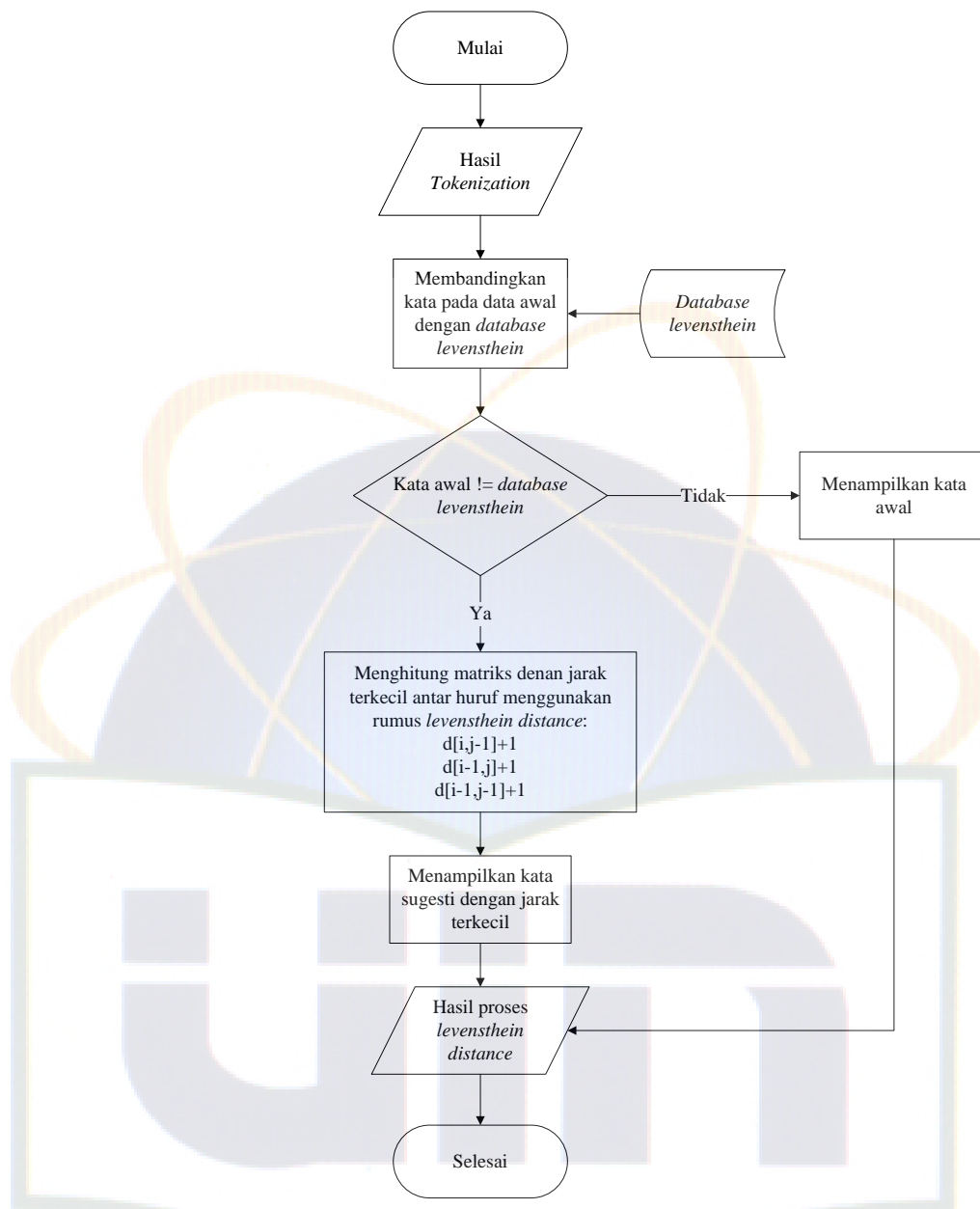
Gambar 4.2 *Flowchart Case Folding*

2. Tahap ketiga yaitu melakukan proses *tokenization*, atau proses pemecahan suatu kalimat menjadi kata-kata.



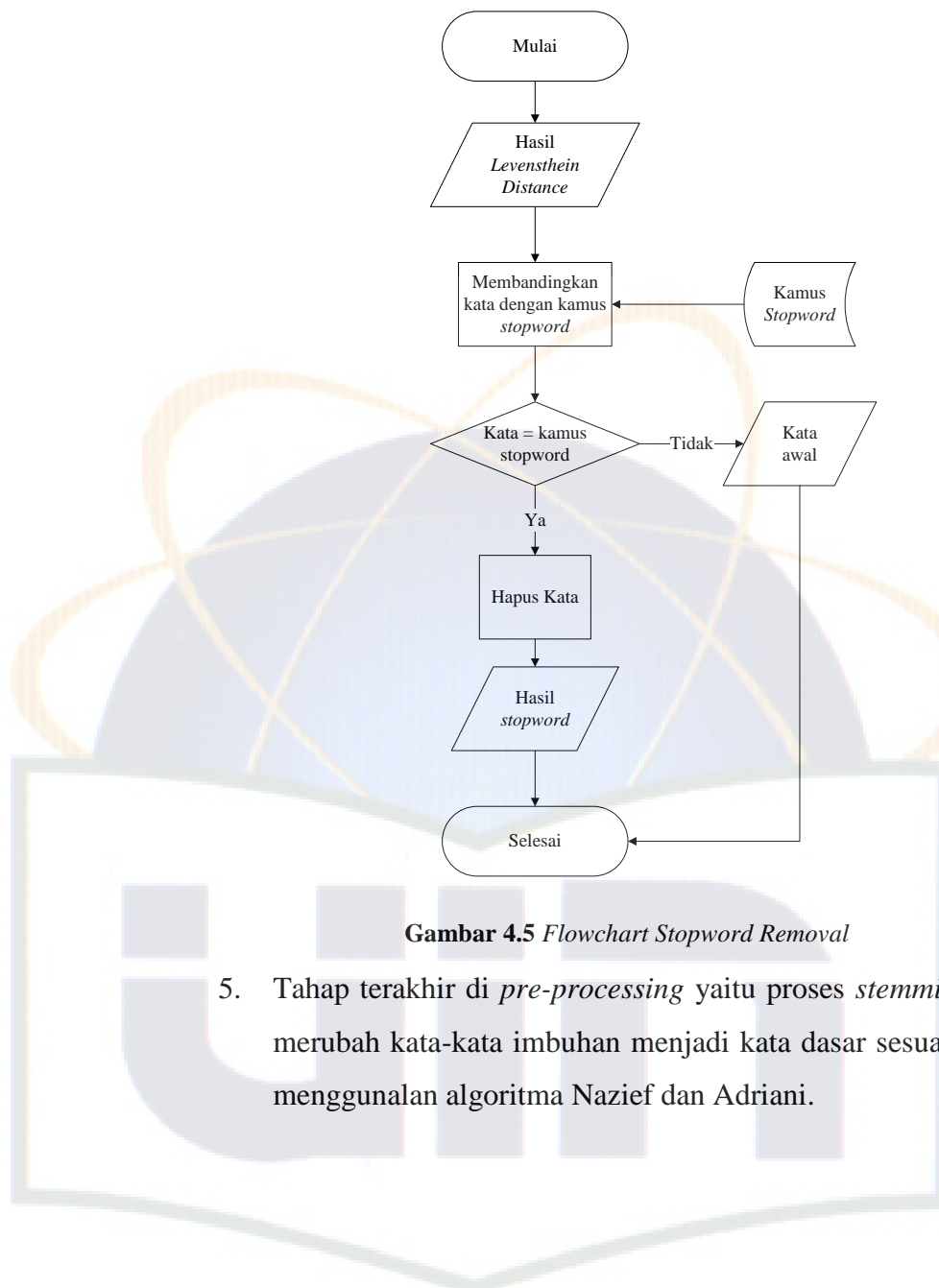
Gambar 4.3 *Flowchart Tokenization*

3. Proses selanjutnya yaitu melakukan proses menormalisasikan sebuah kata yang salah atau *typo* menggunakan algoritma *levenshtein distance* dengan cara menghapus kata yang berlebih, menambah kata yang kurang, dan mengganti kata yang tidak sesuai.



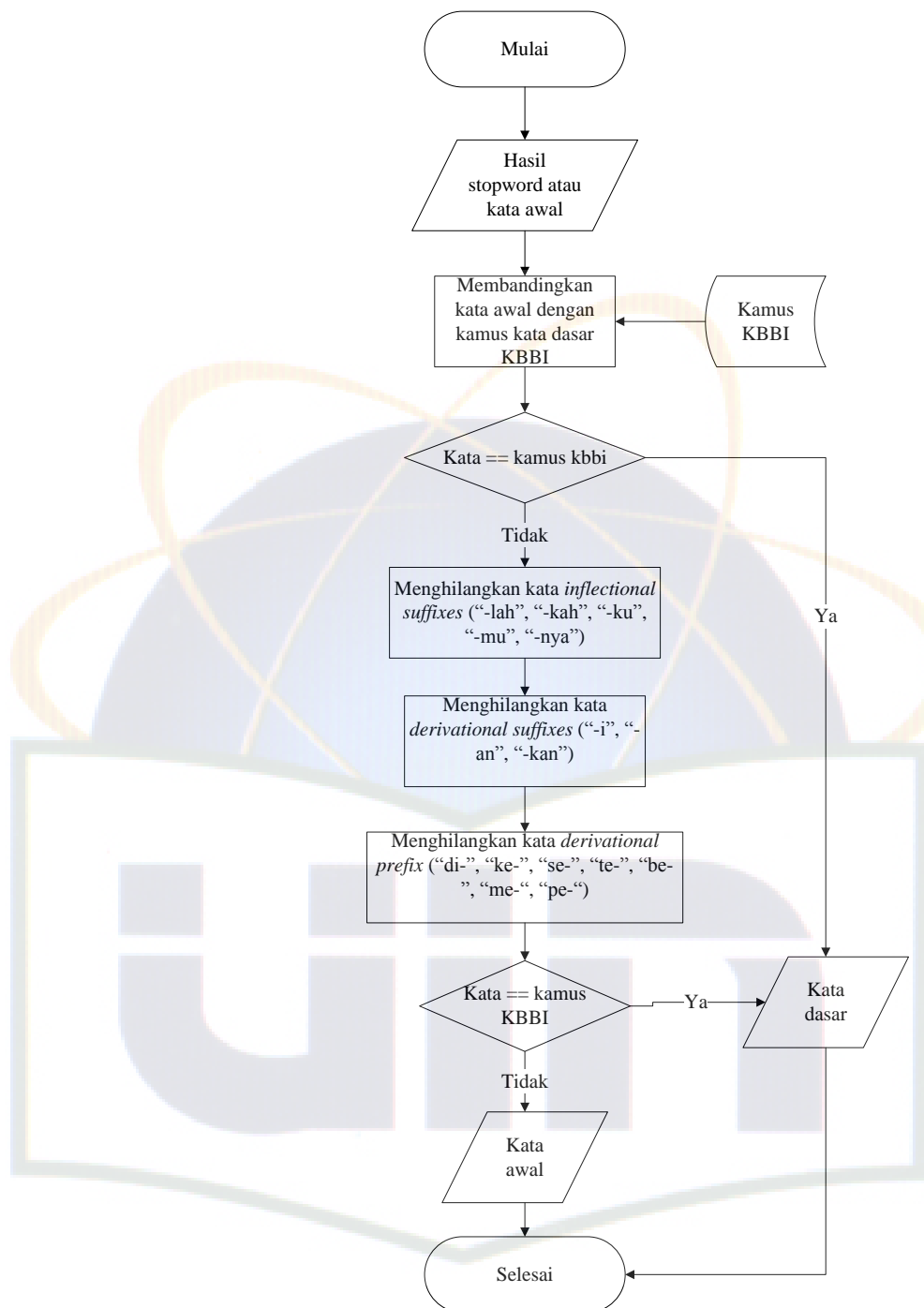
Gambar 4.4 Flowchart Levenshtein Distance

4. Selanjutnya melakukan proses *stopword removal* atau proses penghapusan kata-kata yang dianggap tidak penting, seperti dia, aku, kamu, pada, dan lainnya.



Gambar 4.5 *Flowchart Stopword Removal*

5. Tahap terakhir di *pre-processing* yaitu proses *stemming* atau merubah kata-kata imbuhan menjadi kata dasar sesuai KBBI menggunakan algoritma Nazief dan Adriani.

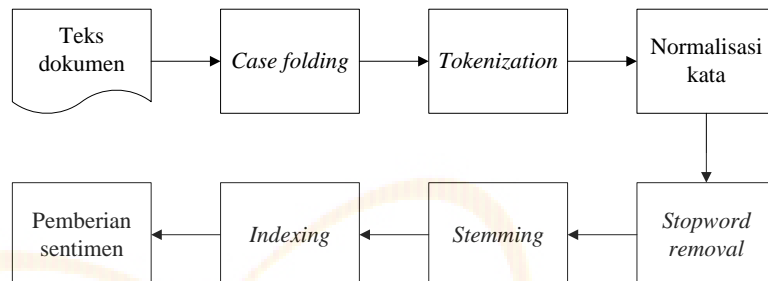


Gambar 4.6 Flowchart Stemming Algoritma Nazief dan Adriani

4.2.2 Conceptual Model Sentimen Pada Data Latih

Konsep penentuan sentimen data latih pada penelitian ini dilakukan secara manual. Setelah diberikan sentimen pada setiap

dokumen data latih, akan dilakukan proses *pre-processing*. Untuk lebih jelasnya lagi, alur pada data latih ini sebagai berikut:



Gambar 4.7 Flowchart penentuan sentimen data latih

Berikut contoh dari setiap proses saat melakukan penentuan sentimen pada data latih:

Teks dokumen:

KEBRSIHAN sebagian dari IMAN !!!

1. Pemberian sentiment

Pada konsep data latih ini, pemberian sentimen akan diberikan secara manual dari setiap dokumen yang dijadikan data latih tersebut. Dimana sentimen yang akan dianalisis pada penelitian ini hanya sentiment positif dan sentiment negatif. Pada contoh ini penulis memberikan sentiment positif.

2. Case folding

Tabel 4.1 Contoh Proses *Case Folding*

Teks dokumen	Proses <i>case folding</i>
KEBRSIHAN sebagian dari IMAN !!!	kebrsihan sebagian dari iman !!!

3. Tokenization

Tabel 4.2 Contoh Proses *Tokenization*

Hasil <i>filtering</i>	Proses <i>tokenization</i>
------------------------	----------------------------

kebrsihan sebagian dari iman	kebrsihan sebagian dari iman
-------------------------------------	---------------------------------------

4. Normalisasi kata (menggunakan algoritma *levensthein distance*)

Tabel 4.3 Contoh Proses Normalisasi Kata

Hasil <i>tokenization</i>	Proses normalisasi
kebrsihan sebagian dari iman	kebersihan sebagian dari iman

5. *Stopword removal*

Tabel 4.4 Contoh Proses *Stopword Removal*

Hasil normalisasi	Proses <i>stopword removal</i>
kebersihan sebagian dari iman	kebersihan iman

6. *Stemming*

Tabel 4.5 Contoh Proses *Stemming*

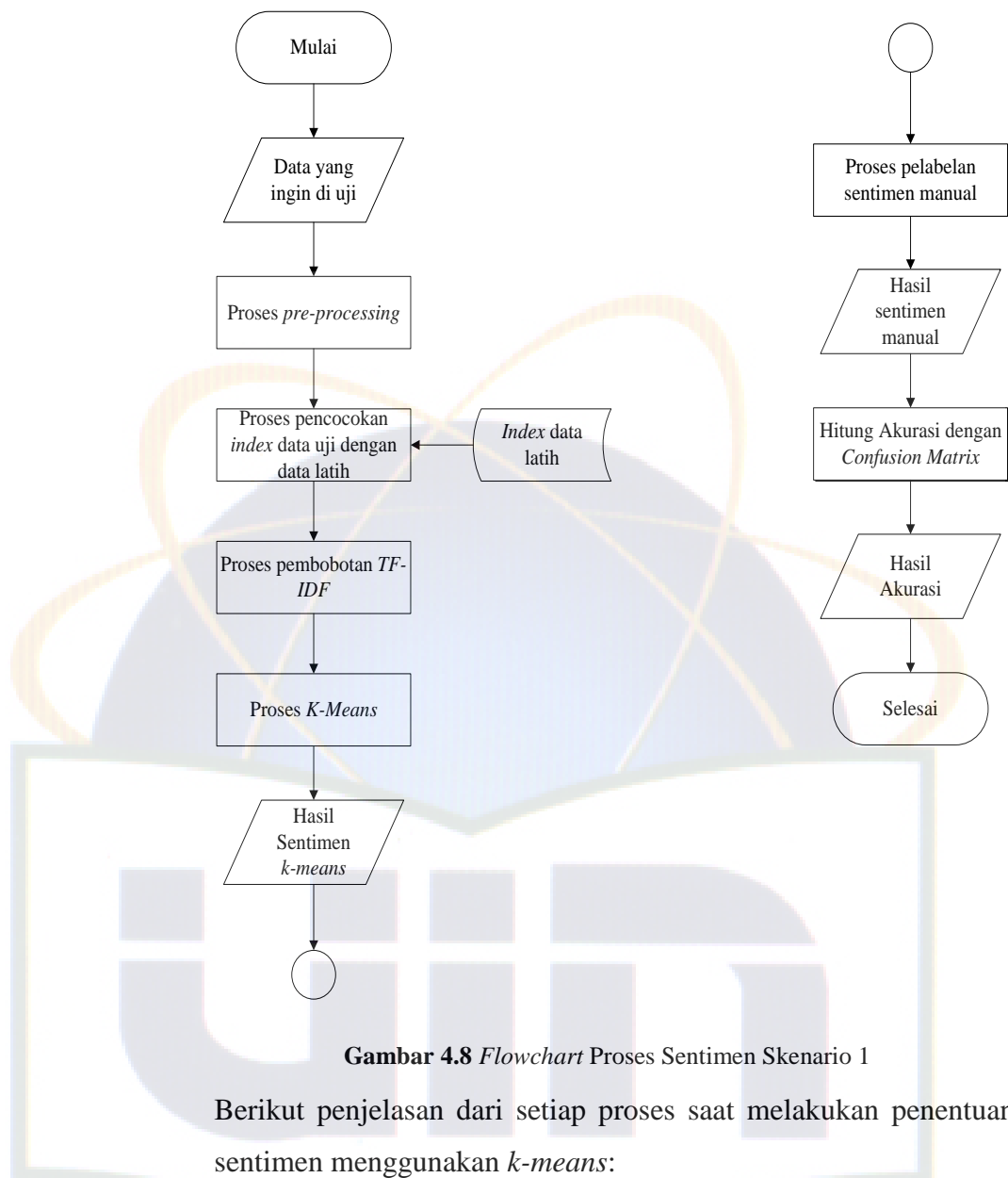
Hasil <i>stopword removal</i>	Proses <i>stemming</i>
kebersihan iman	bersih iman

7. *Indexing*

Pada tahap ini, dilakukan proses pengindeksan pada hasil *pre-processing* dan sentiment dari setiap dokumen. Pada *inverted index* akan tersimpan informasi berupa id dokumen, komentar, hasil *pre-processing* dan hasil sentimen.

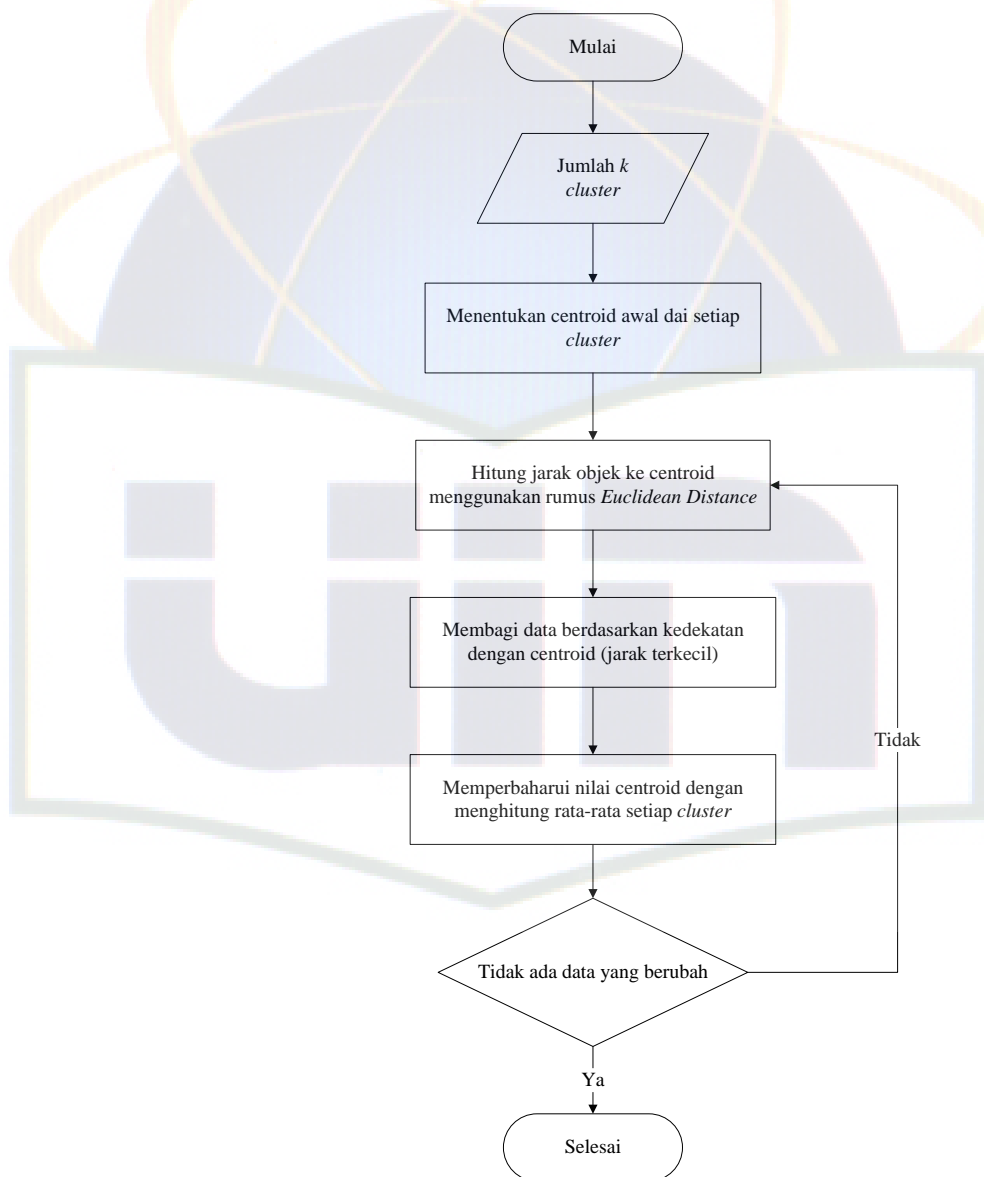
4.2.3 Conceptual Model Sentimen dengan Algoritma *K-Means*

Skenario pertama pada penelitian ini secara alur dari analisa sentimen dengan menggunakan algoritma *k-means* tanpa bantuan algoritma *levensthein distance* sebagai algoritma normalisasi kata dapat dijelaskan pada gambar dibawah ini:



1. Mengumpulkan data yang ingin di uji
2. Melakukan proses *pre-processing* tanpa bantuan algoritma *levenshtein distance* sebagai algoritma normalisasi kata pada data uji
3. Hasil *index* dari *pre-processing* pada data uji, dicari yang cocok atau yang sama pada *index* data latih. Kemudian ambil kalimat yang terdapat *index* yang cocok atau sama itu untuk ke tahap selanjutnya.

4. Lakukan pembobotan kata pada *index* yang ada pada kalimat yang sudah diambil ditahap sebelumnya dan data ujinya menggunakan algoritma *tf-idf* dan diambil nilai total *weighting*nya untuk diproses ditahap selanjutnya.
5. Dari nilai total *weighting* yang sudah didapat dari proses *tf-idf* akan diproses menggunakan algoritma *k-means* untuk menentukan sentimennya. Alur dari algoritma *k-means* dapat digambarkan seperti berikut:



Gambar 4.9 Flowchart Proses Algoritma K-Means

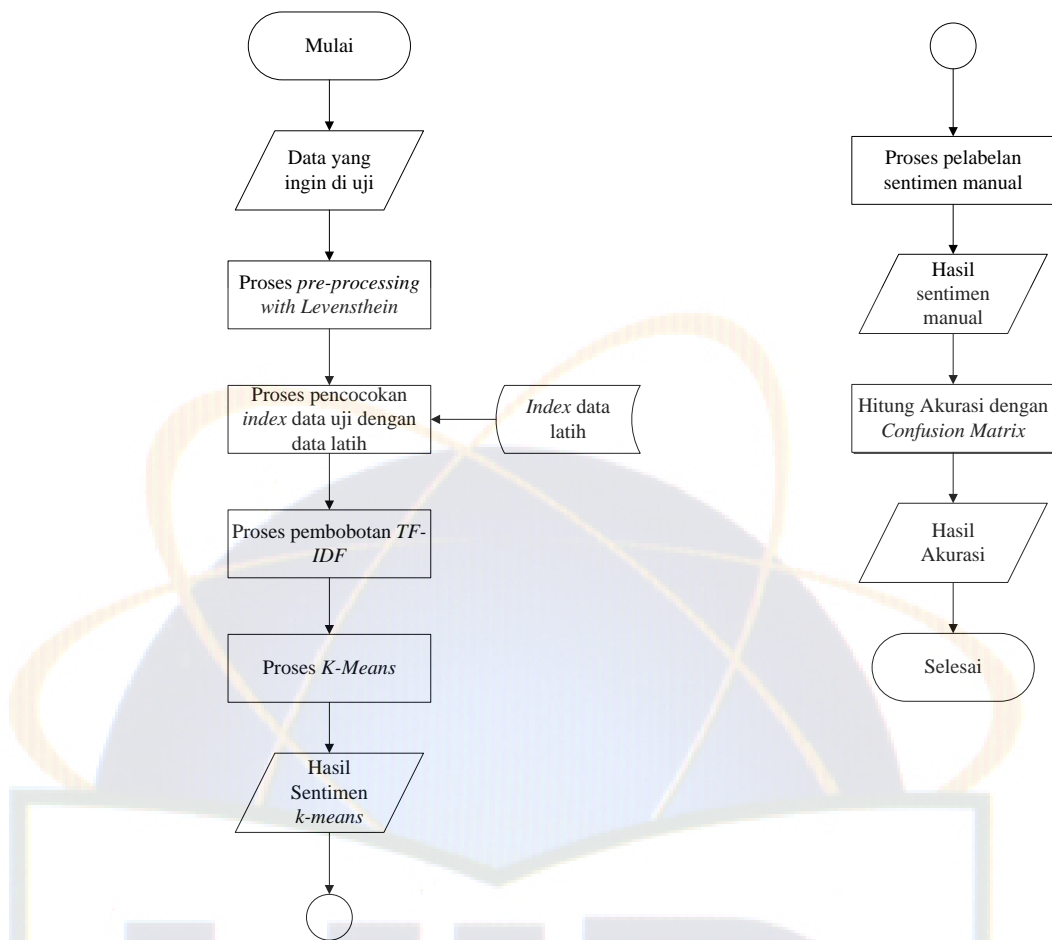
Penjelasan untuk alur algoritma *k-means* sudah dijelaskan pada sub-bab 2.8.2.

Setelah melakukan proses diatas, cari di *cluster* manakah data ujinya berada. Setelah ditemukan, hiraukan *cluster* yang lainnya. Kemudian kita lihat sentiment apakah yang paling banyak pada *cluster* tersebut. Dan itulah hasil sentimen yang diambil.

6. Menentukan pelabelan sentimen secara manual terhadap data uji sebagai perbandingan pada tahap akhir nanti.
7. Hitung akurasi dari perbandingan yang sudah dilakukan sebelumnya antara sentimen yang dihasilkan dari algoritma *k-means* dan sentimen dari pelabelan manual menggunakan rumus *confusion matrix*. Kalau sentimen yang dihasilkan keduanya sama maka akurasinya baik, dan sebaliknya kalau sentimen yang dihasilkan keduanya berbeda maka akurasi tidak baik.
8. Mendapatkan nilai akurasi dari scenario pertama.

4.2.4 Conceptual Model Sentimen Algoritma K-Means dengan Bantuan Algoritma Levensthein Distance

Skenario kedua pada penelitian ini secara alur dari analisa sentimen dengan menggunakan algoritma *k-means* dan bantuan algoritma *levensthein distance* sebagai algoritma normalisasi kata dapat dijelaskan pada gambar dibawah ini:



Gambar 4.10 Flowchart Proses Sentimen Skenario 2

Berikut penjelasan dari setiap proses saat melakukan penentuan sentimen menggunakan *k-means* dan *levensthein distance*:

1. Mengumpulkan data yang ingin di uji
2. Melakukan proses *pre-processing* dengan bantuan algoritma *levensthein distance* sebagai algoritma normalisasi kata pada data uji
3. Hasil *index* dari *pre-processing* pada data uji, dicari yang cocok atau yang sama pada *index* data latih. Kemudian ambil kalimat yang terdapat *index* yang cocok atau sama itu untuk ke tahap selanjutnya.
4. Lakukan pembobotan kata pada *index* yang ada pada kalimat yang sudah diambil ditahap sebelumnya dan data ujinya

menggunakan algoritma *tf-idf* dan diambil nilai total *weighting*nya untuk diproses ditahap selanjutnya.

5. Dari nilai total *weighting* yang sudah didapat dari proses *tf-idf* akan diproses menggunakan algoritma *k-means* untuk menentukan sentimennya. Setelah melakukan proses *k-means*, cari di *cluster* manakah data ujinya berada. Setelah ditemukan, hiraukan *cluster* yang lainnya. Kemudian kita lihat sentiment apakah yang paling banyak pada *cluster* tersebut. Dan itulah hasil sentimen yang kita ambil.
6. Menentukan pelabelan sentimen secara manual terhadap data uji sebagai perbandingan pada tahap akhir nanti.
7. Hitung akurasi dari perbandingan yang sudah dilakukan sebelumnya antara sentimen yang dihasilkan dari algoritma *k-means* dan sentimen dari pelabelan manual menggunakan rumus *confusion matrix*. Kalau sentimen yang dihasilkan keduanya sama maka akurasinya baik, dan sebaliknya kalau sentimen yang dihasilkan keduanya berbeda maka akurasi tidak baik.
8. Mendapatkan nilai akurasi dari scenario kedua.

4.3 Data Masukan/Keluaran (*Input / Output Data*)

Pada penelitian ini data yang diperoleh sebagai data input yaitu komentar-komentar masyarakat terhadap kebijakan pemerintah tentang zonasi sekolah pada peraturan PPDB yaitu sebanyak 200 data yang diambil dari komentar posting *channel* YouTube CNN Indonesia dan komentar *posting* Facebook page Kemendikbud RI. Dimana proses pengambilan data ini dinamakan dengan teknik *crawling*. Pada pengambilan data dari *channel* Youtube peneliti memanfaatkan yang namanya YouTube API dan dari Facebook peneliti mengambilnya secara manual. Selain data-data komentar yang dikumpulkan pada penelitian ini yaitu kamus kata dasar KBBI, *stopwords*, dan kamus *levensthein*.

Hasi atau *output* yang didapatkan dari penelitian ini yaitu sentimen akhir dari suatu komentar dan hasil akurasi pada skenario pertama dan kedua.

4.4 Pemodelan (*Modelling*)

Dalam *modelling phase* atau fase pemodelan pada penelitian ini, dilakukan pemodelan konstruksi analisis sentimen dengan menggunakan algoritma *k-means* saja dan kombinasi algoritma *k-means* dan algoritma *levensthein distance* sebagai algoritma normalisasi kata. Berikut ini dapat dilihat pemodelan-pemodelan tersebut secara lengkap.

4.4.1 Konstruksi Sentimen dengan Algoritma *K-Means*

Konstruksi menggunakan algoritma *k-means* dalam menentukan sentimen ini merupakan salah satu skenario di dalam penelitian saat ini. Dimana dalam penentuan sentimen menggunakan algoritma *k-means* membutuhkan proses penunjang lainnya agar hasilnya lebih maksimal. Secara keseluruhan konstruksi analisa sentimen menggunakan algoritma *k-means* dapat dijelaskan dibawah ini (konsep diambil dari sub-bab 4.2.3 dan dapat dilihat pada gambar 4.8):

1. Pelatihan data latih

Proses dan contoh dalam pelatihan data latih (konsep diambil dari sub-bab 4.2.2 dan dapat dilihat pada gambar 4.7):

- a. Mengumpulkan dokumen yang didapat dari kumpulan komentar yang sudah *dicrawling* sebelumnya. Sebagai contoh digunakan 8 dokumen sebagai data latih.

Tabel 4.6 Dokumen data latih

No.	Komentar
1.	Kemendikbut cerdas mari sikapi kebijakan zonasi PPDB dengan bijak dan pemerintah dlm hal ini kemendikbud

	juga harus siap mengatasi kendala mengenai sistem zonasi tersebut
2.	Tujuannya baik, tp pelaksanaannya sangat rumit, spt daerah kami, sekolah2 banyak yg belum sesuai dgn 8 SNP, oleh sebab itu Kemendikbud dan Pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana
3.	Sistim zonasi tidak efektif...kasian rumahnya yg agak jauh...akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar
4.	InshaaAllah mutu seluruh sekolah di Indonesia bisa sama baik negeri maupun swasta . . itulah salah satu maksud dan tujuannya
5.	Seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi. Faktanya banyak sekolah yg sarana dan prasarananya sangat2 minim terutama ruang kelas yg tidak memadai, akhirnya PBM pun berjalan semrawut
6.	Zonasi demi keadilan sebenarnya tidak adil.
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah, sistem zonasi gak guna.
8.	Sistim zonasi mncecewakan bnyak phak...kpingin sekola yg lbih dekat ja g bsa msuk zonasi

b. Memberikan sentimen secara manual setiap dokumennya

Tabel 4.7 Penentuan sentimen data latih

No.	Komentar	Sentimen
1.	Kemendikbut cerdas mari sikapi kebijakan zonasi PPDB dengan bijak dan pemerintah dlm hal ini kemendikbud juga	Positif

	harus siap mengatasi kendala mengenai sistem zonasi tersebut	
2.	Tujuannya baik, tp pelaksanaannya sangat rumit, spt daerah kami, sekolah2 banyak yg belum sesuai dgn 8 SNP, oleh sebab itu Kemendikbud dan Pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana	Negatif
3.	Sistim zonasi tidak efektif...kasion rumahnya yg agak jauh...akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar	Negatif
4.	InshaaAllah mutu seluruh sekolah di Indonesia bisa sama baik negeri maupun swasta . . itulah salah satu maksud dan tujuannya	Positif
5.	Seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi. Faktanya banyak sekolah yg sarana dan prasarannya sangat2 minim terutama ruang kelas yg tidak memadai, akhirnya PBM pun berjalan semrawut	Negatif
6.	Zonasi demi keadilan sebenarnya tidak adil.	Negatif
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah, sistem zonasi gak guna.	Negatif
8.	Sistim zonasi mngecewakan banyak pihak...kpingin sekola yg lebih dekat ja g bsa masuk zonasi	Negatif

c. Melakukan proses *case folding***Tabel 4.8** Hasil Proses *Case Folding*

No.	Komentar
1.	kemendikbut cerdas mari sikapi kebijakan zonasi ppdb dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut
2.	tujuannya baik, tp pelaksanaannya sangat rumit, spt daerah kami, sekolah2 banyak yg belum sesuai dgn 8 snp, oleh sebab itu kemendikbud dan pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana
3.	sistim zonasi tidak efektif...kasian rumahnya yg agak jauh...akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar
4.	inshaaallah mutu seluruh sekolah di indonesia bisa sama baik negeri maupun swasta . . itulah salah satu maksud dan tujuannya
5.	seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi. faktanya banyak sekolah yg sarana dan prasarananya sangat2 minim terutama ruang kelas yg tidak memadai, akhirnya pbm pun berjalan semrawut
6.	zonasi demi keadilan sebenarnya tidak adil.
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah, sistem zonasi gak guna.
8.	sistim zonasi mngecewakan bnyak phak...kpingin sekola yg lebih dekat ja g bsa masuk zonasi

d. Melakukan proses *filtering***Tabel 4.9** Hasil Proses *Filtering*

No.	Komentar
1.	kemendikbut cerdas mari sikapi kebijakan zonasi ppdb dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut
2.	tujuannya baik tp pelaksanaannya sangat rumit spt daerah kami sekolah banyak yg belum sesuai dgn smp oleh sebab itu kemendikbud dan pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana
3.	sistim zonasi tidak efektif kesian rumahnya yg agak jauh akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar
4.	inshaaallah mutu seluruh sekolah di indonesia bisa sama baik negeri maupun swasta itulah salah satu maksud dan tujuannya
5.	seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi faktanya banyak sekolah yg sarana dan prasarananya sangat minim terutama ruang kelas yg tidak memadai akhirnya pbm pun berjalan semrawut
6.	zonasi demi keadilan sebenarnya tidak adil
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah sistem zonasi gak guna
8.	sistim zonasi mngecewakan banyak pihak kpingin sekola yg lebih dekat ja g bsa masuk zonasi

e. Melakukan proses *tokenization*

Tabel 4.10 Hasil Proses *Tokenization*

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
kemendikbut cerdas	tujuannya baik	sistim zonasi	inshaaallah mutu

mari sikapi kebijakan zonasi ppdb dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut	tp pelaksanaannya sangat rumit spt daerah kami sekolah banyak yg belum sesuai dgn snp oleh sebab itu kemendikbud dan pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana	tidak efektif kasian rumahnya yg agak jauh akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar	seluruh sekolah di indonesia bisa sama baik negeri maupun swasta itulah salah satu maksud dan tujuannya
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
seharusnya	zonasi	selama	sistim

mendikbut	demi	sarana	zonasi
kaji	keadilan	prasarana	mngecewakan
dulu	sebenarnya	dan	banyak
dilapangan	tidak	sdm	pihak
sebelum	adil	masih	kpingin
menerapkan		belum	sekola
sisten		merata	yg
zonasi		di	lebih
faktanya		tiap	dekat
banyak		sekolah	ja
sekolah		sistem	g
yg		zonasi	bisa
sarana		gak	masuk
dan		guna	zonasi
prasarananya			
sangat			
minim			
terutama			
ruang			
kelas			
yg			
tidak			
memadai			
akhirnya			
pbm			
pun			
berjalan			
semrawut			

f. Melakukan proses normalisasi kata

Tabel 4.11 Hasil Proses Normalisasi Kata

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
kemendikbud	tujuannya	sistem	insya allah
cerdas	baik	zonasi	mutu
mari	tapi	tidak	seluruh
sikapi	pelaksanaannya	efektif	sekolah
kebijakan	sangat	kasihan	di
zonasi	rumit	rumahnya	indonesia
ppdb	seperti	yang	bisa
dengan	daerah	agak	sama
bijak	kami	jauh	baik
dan	sekolah	akhirnya	negeri
pemerintah	banyak	mereka	maupun
dalam	yang	dengan	swasta
hal	belum	berat	itulah
ini	sesuai	hati	salah
kemendikbud	dengan	masukin	satu
juga	snp	anaknya	maksud
harus	oleh	ke	dan
siap	sebab	sekolah	tujuannya
mengatasi	itu	swasta	
kendala	kemendikbud	yang	
mengenai	dan	serba	
sistem	pemerintah	bayar	
zonasi	daerah		
tersebut	harus		
	berani		
	membuat		
	terobosan		
	jangan		
	sekedar		
	wacana		

Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
seharusnya kemendikbud kaji dahulu dilapangan sebelum menerapkan sistem zonasi faktanya banyak sekolah yang sarana dan prasarananya sangat minim terutama ruang kelas yang tidak memadai akhirnya pbm pun berjalan semrawut	zonasi demi keadilan sebenarnya tidak adil	selama sarana prasarana dan sdm masih belum merata di tiap sekolah sistem zonasi tidak guna	sistem zonasi mengecewakan banyak pihak kepingin sekolah yang lebih dekat saja tidak bisa masuk zonasi

g. Melakukan proses *stemming*

Tabel 4.12 Hasil Proses *Stemming*

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
kemendikbud	tuju	sistem	insya allah
cerdas	baik	zonasi	mutu
mari	tapi	tidak	seluruh
sikap	laksana	efektif	sekolah
bijak	sangat	kasihan	di
zonasi	rumit	rumah	indonesia
ppdb	seperti	yang	bisa
dengan	daerah	agak	sama
bijak	kami	jauh	baik
dan	sekolah	akhir	negeri
pemerintah	banyak	mereka	maupun
dalam	yang	dengan	swasta
hal	belum	berat	itu
ini	sesuai	hati	salah
kemendikbud	dengan	masuk	satu
juga	snp	anak	maksud
harus	oleh	ke	dan
siap	sebab	sekolah	tuju
atas	itu	swasta	
kendala	kemendikbud	yang	
kena	dan	serba	
sistem	pemerintah	bayar	
zonasi	daerah		
sebut	harus		
	berani		
	buat		
	terobos		
	jangan		

	sekedar wacana		
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
<p>harus kemendikbud kaji dahulu lapang belum terap sistem zonasi fakta banyak sekolah yang sarana dan prasarana sangat minim utama ruang kelas yang tidak ada akhir pbm pun</p>	<p>zonasi demi adil benar tidak adil</p>	<p>selama sarana prasarana dan sdm masih belum rata di tiap sekolah sistem zonasi tidak guna</p>	<p>sistem zonasi kecewa banyak pihak kepingin sekolah yang lebih dekat saja tidak bisa masuk zonasi</p>

jalan			
semrawut			

h. Melakukan proses *filtering* / *stopword*

Tabel 4.13 Hasil Proses *Filtering* / *Stopword*

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
cerdas mari sikap bijak zonasi bijak pemerintah harus siap atas kendala kena sistem zonasi sebut	tuju baik laksana rumit daerah sekolah banyak belum sesuai sebab pemerintah daerah harus berani buat terobos sekedar wacana	sistem zonasi tidak efektif kasihan rumah jauh akhir berat hati masuk anak sekolah swasta serba bayar	insya allah mutu seluruh sekolah indonesia bisa sama baik negeri swasta salah maksud tuju
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
harus kaji dahulu lapang belum terap	zonasi demi adil benar tidak adil	selama sarana prasarana sdm masih belum	sistem zonasi kecewa banyak pihak kepingin

sistem		rata	sekolah
zonasi		tiap	lebih
fakta		sekolah	dekat
banyak		sistem	tidak
sekolah		zonasi	bisa
sarana		tidak	masuk
prasarana		guna	zonasi
minim			
utama			
ruang			
kelas			
tidak			
ada			
akhir			
jalan			
semrawut			

- i. Dari proses *pre-processing* pada data latih, hasilnya dibuat *index* untuk disimpan oleh variable.
2. Menyiapkan data yang ingin diuji. Sebagai contoh akan dilakukan 1 data uji agar prosesnya terlihat.

Tidak efektif karena belum semuanya sekolah di Indonesia bagus

3. Melakukan *pre-processing* pada data uji. Pada tahap awal akan melakukan proses *case folding*.

tidak efektif karena belum semuanya sekolah di indonesia bagus

4. Melakukan proses *tokenization*

tidaak
efektif

karena
belum
semuanya
sekolah
di
indonesia
bagus

5. Melakukan proses normalisasi kata ejaan tanpa bantuan algoritma *levensthein distance* untuk *typo*. Pada contoh ini tidak ada yang di normalisasi kata ejaan, karena tidak ada kata ejaan didalamnya.
6. Melakukan proses *stemming*, merubah kata berimbuhan mejadi kata dasar sesuai dengan KBBI

tidaak
efejtif
karena
belum
semua
sekolah
di
indonesia
bagus

7. Melakukan proses *filtering / stopword*.

tidaak
efejtif
karena
belum
semua
sekolah

indonesia
bagus

8. Melakukan tahap terakhir di *pre-processing* pada penelitian ini yaitu *indexing* pada data uji.
9. Melakukan pencarian atau pencocokan *index* yang ada di dalam data latih dan *index* yang ada di data uji. Didapatkan *index* data uji yang sama dengan data latih yaitu data ke 2, 3, 4, 5, 7, dan 8 pada data latih.
10. Melakukan proses pembobotan menggunakan algoritma *TF-IDF*. Proses perhitungan awal yaitu menghitung *IDF* dengan rumus yang dapat dilihat pada rumus 2.7.

Tabel 4.14 Hasil Perhitungan *IDF* Skenario 1

	IDF	TF-0	TF-1	TF-2	TF-3	TF-4	TF-5	TF-6	DF
tidaak	0	1	0	0	0	0	0	0	0
efejtif	0	1	0	0	0	0	0	0	0
karena	0	1	0	0	0	0	0	0	0
belum	0.176091	1	1	0	0	1	0	1	4
semua	0.778151	1	0	0	0	0	0	0	1
sekolah	0	1	1	1	1	1	1	1	6
indonesia	0	1	0	0	0	0	0	0	0
baik	0.477121	1	1	0	1	0	0	0	2
tuju	0.477121	0	1	0	1	0	0	0	2
laksana	0.778151	0	1	0	0	0	0	0	1
rumit	0.778151	0	1	0	0	0	0	0	1
daerah	0.778151	0	2	0	0	0	0	0	1
banyak	0.30103	0	1	0	0	1	0	1	3
sesuai	0.778151	0	1	0	0	0	0	0	1
sebab	0.778151	0	1	0	0	0	0	0	1
pemerintah	0.778151	0	1	0	0	0	0	0	1
harus	0.477121	0	1	0	0	1	0	0	2
berani	0.778151	0	1	0	0	0	0	0	1
buat	0.778151	0	1	0	0	0	0	0	1
terobos	0.778151	0	1	0	0	0	0	0	1
sekedar	0.778151	0	1	0	0	0	0	0	1
wacana	0.778151	0	1	0	0	0	0	0	1
sistem	0.176091	0	0	1	0	1	1	1	4

zonasi	0.176091	0	0	1	0	1	1	1	4
tidak	0.176091	0	0	1	0	1	1	1	4
efektif	0.778151	0	0	1	0	0	0	0	1
kasihan	0.778151	0	0	1	0	0	0	0	1
rumah	0.778151	0	0	1	0	0	0	0	1
jauh	0.778151	0	0	1	0	0	0	0	1
akhir	0.477121	0	0	1	0	1	0	0	2
berat	0.778151	0	0	1	0	0	0	0	1
hati	0.778151	0	0	1	0	0	0	0	1
masuk	0.477121	0	0	1	0	0	0	1	2
anak	0.778151	0	0	1	0	0	0	0	1
swasta	0.477121	0	0	1	1	0	0	0	2
serba	0.778151	0	0	1	0	0	0	0	1
bayar	0.778151	0	0	1	0	0	0	0	1
insya allah	0.778151	0	0	0	1	0	0	0	1
mutu	0.778151	0	0	0	1	0	0	0	1
seluruh	0.778151	0	0	0	1	0	0	0	1
bisa	0.477121	0	0	0	1	0	0	1	2
sama	0.778151	0	0	0	1	0	0	0	1
negeri	0.778151	0	0	0	1	0	0	0	1
salah	0.778151	0	0	0	1	0	0	0	1
maksud	0.778151	0	0	0	1	0	0	0	1
kaji	0.778151	0	0	0	0	1	0	0	1
dahulu	0.778151	0	0	0	0	1	0	0	1
lapang	0.778151	0	0	0	0	1	0	0	1
terap	0.778151	0	0	0	0	1	0	0	1
fakta	0.778151	0	0	0	0	1	0	0	1
sarana	0.477121	0	0	0	0	1	1	0	2
prasarana	0.477121	0	0	0	0	1	1	0	2
minim	0.778151	0	0	0	0	1	0	0	1
utama	0.778151	0	0	0	0	1	0	0	1
ruang	0.778151	0	0	0	0	1	0	0	1
kelas	0.778151	0	0	0	0	1	0	0	1
ada	0.778151	0	0	0	0	1	0	0	1
jalan	0.778151	0	0	0	0	1	0	0	1
semrawut	0.778151	0	0	0	0	1	0	0	1
selama	0.778151	0	0	0	0	0	1	0	1
sdm	0.778151	0	0	0	0	0	1	0	1
masih	0.778151	0	0	0	0	0	1	0	1
rata	0.778151	0	0	0	0	0	1	0	1
tiap	0.778151	0	0	0	0	0	1	0	1

guna	0.778151	0	0	0	0	0	1	0	1
kecewa	0.778151	0	0	0	0	0	1	0	1
pihak	0.778151	0	0	0	0	0	1	0	1
kepingin	0.778151	0	0	0	0	0	1	0	1
lebih	0.778151	0	0	0	0	0	1	0	1
dekat	0.778151	0	0	0	0	0	1	0	1

Setelah mendapatkan nilai *IDF*, akan menghitung nilai bobotnya (*term weighting*) dengan rumus yang dapat dilihat pada rumus 2.8.

Tabel 4.15 Hasil Perhitungan bobot *TF-IDF* Skenario 2

	W-0	W-1	W-2	W-3	W-4	W-5	W-6
tidaak	0	0	0	0	0	0	0
efektif	0	0	0	0	0	0	0
karena	0	0	0	0	0	0	0
belum	0.176091	0.176091	0	0	0.176091	0	0.176091
semua	0.778151	0	0	0	0	0	0
sekolah	0	0	0	0	0	0	0
indonesia	0	0	0	0	0	0	0
baik	0.477121	0.477121	0	0.477121	0	0	0
tuju	0	0.477121	0	0.477121	0	0	0
laksana	0	0.778151	0	0	0	0	0
rumit	0	0.778151	0	0	0	0	0
daerah	0	1.556303	0	0	0	0	0
banyak	0	0.30103	0	0	0.30103	0	0.30103
sesuai	0	0.778151	0	0	0	0	0
sebab	0	0.778151	0	0	0	0	0
pemerintah	0	0.778151	0	0	0	0	0
harus	0	0.477121	0	0	0.477121	0	0
berani	0	0.778151	0	0	0	0	0
buat	0	0.778151	0	0	0	0	0
terobos	0	0.778151	0	0	0	0	0
sekedar	0	0.778151	0	0	0	0	0
wacana	0	0.778151	0	0	0	0	0
sistem	0	0	0.176091	0	0.176091	0.176091	0.176091
zonasi	0	0	0.176091	0	0.176091	0.176091	0.176091
tidak	0	0	0.176091	0	0.176091	0.176091	0.176091
efektif	0	0	0.778151	0	0	0	0
kasihan	0	0	0.778151	0	0	0	0
rumah	0	0	0.778151	0	0	0	0

jauh	0	0	0.778151	0	0	0	0
akhir	0	0	0.477121	0	0.477121	0	0
berat	0	0	0.778151	0	0	0	0
hati	0	0	0.778151	0	0	0	0
masuk	0	0	0.477121	0	0	0	0.477121
anak	0	0	0.778151	0	0	0	0
swasta	0	0	0.477121	0.477121	0	0	0
serba	0	0	0.778151	0	0	0	0
bayar	0	0	0.778151	0	0	0	0
insya allah	0	0	0	0.778151	0	0	0
mutu	0	0	0	0.778151	0	0	0
seluruh	0	0	0	0.778151	0	0	0
bisa	0	0	0	0.477121	0	0	0.477121
sama	0	0	0	0.778151	0	0	0
negeri	0	0	0	0.778151	0	0	0
salah	0	0	0	0.778151	0	0	0
maksud	0	0	0	0.778151	0	0	0
kaji	0	0	0	0	0.778151	0	0
dahulu	0	0	0	0	0.778151	0	0
lapang	0	0	0	0	0.778151	0	0
terap	0	0	0	0	0.778151	0	0
fakta	0	0	0	0	0.778151	0	0
sarana	0	0	0	0	0.477121	0.477121	0
prasarana	0	0	0	0	0.477121	0.477121	0
minim	0	0	0	0	0.778151	0	0
utama	0	0	0	0	0.778151	0	0
ruang	0	0	0	0	0.778151	0	0
kelas	0	0	0	0	0.778151	0	0
ada	0	0	0	0	0.778151	0	0
jalan	0	0	0	0	0.778151	0	0
semrawut	0	0	0	0	0.778151	0	0
selama	0	0	0	0	0	0.778151	0
sdm	0	0	0	0	0	0.778151	0
masih	0	0	0	0	0	0.778151	0
rata	0	0	0	0	0	0.778151	0
tiap	0	0	0	0	0	0.778151	0
guna	0	0	0	0	0	0.778151	0
kecewa	0	0	0	0	0	0.778151	0
pihak	0	0	0	0	0	0.778151	0
kepingin	0	0	0	0	0	0.778151	0
lebih	0	0	0	0	0	0.778151	0

dekat	0	0	0	0	0	0.778151	0
TOTAL	1.431364	11.2463	8.962999	7.355544	12.2517	10.04218	1.959638

Dimana total *term weighing* yang didapat pada perhitungan diatas yaitu:

Data uji = 1.431364

Kalimat – 1 = 11.2463

Kalimat – 2 = 8.962999

Kalimat – 3 = 7.355544

Kalimat – 4 = 12.2517

Kalimat – 5 = 10.04218

Kalimat – 6 = 1.959638

11. Melakukan proses *clustering* menggunakan algoritma *k-means* dimana prosesnya dapat dilihat pada gambar 4.9 dan penjelasannya bisa dilihat pada sub-bab 2.8.2. Data yang di *cluster* yaitu jumlah *Term. Weighting* dari setiap data yang sudah dihitung pada proses pembobotan.

- a. Menentukan jumlah *cluster*. Dipenelitian ini menggunakan jumlah *cluster* = 2. Karena *output* yang dihasilkan hanya sentimen positif dan negatif.

cluster 1

cluster 2

--	--

- b. Menentukan centroid awal secara random. Dimana dipenelitian ini centroid awal dari setiap *cluster* yaitu menggunakan nilai terkecil dan nilai terbesar dari jumlah total *term weighing*. Centroid pada

cluster 1 yaitu 1,431364 (data uji) dan centroid pada
cluster 2 yaitu 12,2517 (kalimat – 4)

<i>cluster 1</i>	<i>cluster 2</i>
1,431364 (c.a)	12,2517 (c.a)

Keterangan :

c.a = centroid awal

- c. Menghitung jarak objek ke centroid menggunakan *euclidean distance* untuk mendapatkan jarak terdekat (rumus bisa dilihat pada rumus 2.1).

Tabel 4.16 Perhitungan *Euclidean Distance cluster 1*

	<i>Cluster 1</i>
Data uji	$\sqrt{(1.431364 - 1.431364)^2} = 0$
Kalimat – 1	$\sqrt{(11.2463 - 1.431364)^2} = 9.8149$
Kalimat – 2	$\sqrt{(8.962999 - 1.431364)^2} = 7.5316$
Kalimat – 3	$\sqrt{(7.355544 - 1.431364)^2} = 5.9241$
Kalimat – 4	$\sqrt{(12.2517 - 1.431364)^2} = 10.8203$
Kalimat – 5	$\sqrt{(10.04218 - 1.431364)^2} = 8.6108$
Kalimat – 6	$\sqrt{(1.959638 - 1.431364)^2} = 0.5282$

Tabel 4.17 Perhitungan *Euclidean Distance cluster 2*

	<i>Cluster 2</i>
Data uji	$\sqrt{(1.431364 - 12.2517)^2} = 10.8203$
Kalimat – 1	$\sqrt{(11.2463 - 12.2517)^2} = 1.0054$
Kalimat – 2	$\sqrt{(8.962999 - 12.2517)^2} = 3.2887$
Kalimat – 3	$\sqrt{(7.355544 - 12.2517)^2} = 4.8961$
Kalimat – 4	$\sqrt{(12.2517 - 12.2517)^2} = 0$
Kalimat – 5	$\sqrt{(10.04218 - 12.2517)^2} = 2.2095$
Kalimat – 6	$\sqrt{(1.959638 - 12.2517)^2} = 10.2920$

- d. Membagi semua objek ke *cluster*, jika jarak objek dengan centroid pada *cluster 1* lebih kecil dibanding jarak objek dengan centroid pada *cluster 2* maka objek tersebut masuk ke dalam *cluster 1*, dan sebaliknya.

<i>cluster 1</i>	<i>cluster 2</i>
1,431364 (c.a)	12,2517 (c.a)
1.431364 (data uji)	11.2463 (kal-1)
1.959638 (kal-6)	8.962999 (kal-2)
	7.355544 (kal-3)
	12.2517 (kal-4)
	10.04218 (kal-5)

Didapatkan pembagian *cluster* sesuai dengan perhitungan dari *euclidean distance* setiap objek sebelumnya yaitu seperti yang diatas.

- e. Mencari centroid baru dengan cara menghitung rata – rata nilai dari setiap *cluster*.

Centroid baru *cluster 1* :

$$(1.431364 + 1.959638) / 2 = 1.6955$$

Centroid baru *cluster 2* :

$$(11.2463 + 8.962999 + 7.355544 + 12.2517 + 10.04218) / 5 = 9.9717$$

<i>cluster 1</i>	<i>cluster 2</i>
1.6955 (c.b)	9.9717 (c.b)

Keterangan :

c.b = centroid baru

- f. Menghitung kembali jarak antara objek dengan centroid baru dengan menggunakan *euclidean distance*.

Tabel 4.18 Perhitungan kembali *Euclidean Distance cluster 1*

	<i>Cluster 1</i>
Data uji	$\sqrt{(1.431364 - 1.6955)^2} =$ 0.2641
Kalimat – 1	$\sqrt{(11.2463 - 1.6955)^2} =$ 9.5508
Kalimat – 2	$\sqrt{(8.962999 - 1.6955)^2} =$ 7.2674
Kalimat – 3	$\sqrt{(7.355544 - 1.6955)^2} =$ 5.6600

Kalimat – 4	$\sqrt{(12.2517 - 1.6955)^2} =$ 10.5562
Kalimat – 5	$\sqrt{(10.04218 - 1.6955)^2} =$ 8.3466
Kalimat – 6	$\sqrt{(1.959638 - 1.6955)^2} =$ 0.2641

Tabel 4.19 Perhitungan kembali *Euclidean Distance cluster 2*

	<i>Cluster 2</i>
Data uji	$\sqrt{(1.431364 - 9.9717)^2} =$ 8.5403
Kalimat – 1	$\sqrt{(11.2463 - 9.9717)^2} =$ 1.2746
Kalimat – 2	$\sqrt{(8.962999 - 9.9717)^2} =$ 1.0087
Kalimat – 3	$\sqrt{(7.355544 - 9.9717)^2} =$ 2.6161
Kalimat – 4	$\sqrt{(12.2517 - 9.9717)^2} =$ 2.28
Kalimat – 5	$\sqrt{(10.04218 - 9.9717)^2} =$ 0.3251
Kalimat – 6	$\sqrt{(1.959638 - 9.9717)^2} =$ 8.0120

- g. Membagi semua objek ke *cluster*, sama seperti sebelumnya jika jarak objek dengan centroid pada *cluster 1* lebih kecil dibanding jarak objek dengan centroid pada *cluster 2* maka objek tersebut masuk ke dalam *cluster 1*, dan sebaliknya. Disini akan dilihat apakah hasil pembagiannya terjadi perubahan atau tidak dari pembagian sebelumnya.

<i>cluster 1</i>	<i>cluster 2</i>
1.6074 (c.b)	10.3517 (c.b)
1.431364 (data uji)	11.2463 (kal-1)
1.959638 (kal-6)	8.962999 (kal-2)
	7.355544 (kal-3)
	12.2517 (kal-4)
	10.04218 (kal-5)

Dari hasil pembagian dengan menghitung jarak antara objek dengan centroid baru, tidak terjadi perubahan *cluster* atau objeknya tidak ada yang berubah dari pembagian yang awal.

- h. Karena tidak terjadi perubahan atau perpindahan *cluster* dari setiap objeknya, maka iterasi dari proses *K-Means* berhenti. Selesai.

12. Mencari di *cluster* mana posisi objek dari data uji berada. Apakah di *cluster 1* atau di *cluster 2*. Jika posisi objek dari data uji berada di *cluster 1*, maka *cluster 2* dibiarkan atau tidak digunakan. Dan sebaliknya.

<i>cluster 1</i>	<i>cluster 2</i>
1.6074 (c.b)	10.3517 (c.b)
1.431364 (data uji)	11.2463 (kal-1)
1.959638 (kal-6)	8.962999 (kal-2)
	7.355544 (kal-3)
	12.2517 (kal-4)
	10.04218 (kal-5)

Ditemukan objek dari data uji berada di *cluster 1*, maka *cluster 2* tidak digunakan dan akan fokus pada *cluster 1* saja.

13. Mencari sentimennya yaitu dengan cara dilihat apakah objek kalimat – n pada *cluster* tersebut lebih banyak

bersentimen positif atau negatif. Jika lebih banyak positif, maka sentimen dari data uji yaitu positif, dan sebaliknya. Namun jika jumlah sentimen positif dan negatifnya sama, maka hasil sentimennya yaitu netral.

cluster 1

1.6074 (c.b)
1.431364 (data uji)
1.959638 (kal-6)

Dilihat dari anggota atau objek yang berada di *cluster 1* hanya ada 2 yaitu data uji (yang ingin dicari sentimennya) dan objek kalimat – 6 yang bersentimen negatif. Maka data uji yang dicari sentimennya juga bersentimen negatif.

14. Proses selanjutnya yaitu pelabelan secara manual pada data uji. Dilihat dari susunan katanya, data uji ini diberi sentimen negatif.
15. Menghitung akurasi antara hasil sentimen menggunakan algoritma *K-Means* dan hasil sentimen secara manual menggunakan *confusion matrix*. Rumusnya bisa dilihat pada rumus 2.9.

	<i>Positive</i> (+)	<i>Negative</i> (-)
<i>Positive</i> (+)	0	0
<i>Negative</i> (-)	0	1

$$\frac{\text{Jumlah data yang diprediksi benar}}{\text{Total data yang diprediksi}} = \frac{(0+1)}{(0+0+1+0)} \times 100\% = 100\%$$

16. Selesai.

4.4.2 Konstruksi Sentimen Algoritma *K-Means* dengan Bantuan Algoritma *Levensthein Distance*

Konstruksi menggunakan algoritma *K-Means* dengan bantuan algoritma *Levensthein Distance* dalam menentukan sentimen ini merupakan skenario ke – 2 dipenelitian ini. Dimana dalam penentuan sentimen menggunakan algoritma *K-Means* dengan bantuan algoritma *Levensthein Distance* membutuhkan proses penunjang lainnya agar hasilnya lebih maksimal. Secara keseluruhan konstruksi analisa sentimen menggunakan algoritma *K-Means* dengan bantuan algoritma *Levensthein Distance* dapat dijelaskan dibawah ini (konsep diambil dari sub-bab 4.2.4 dan dapat dilihat pada gambar 4.10):

1. Pelatihan data latih

Proses dan contoh dalam pelatihan data latih (konsep diambil dari sub-bab 4.2.2 dan dapat dilihat pada gambar 4.7):

- a. Mengumpulkan dokumen yang didapat dari kumpulan komentar yang sudah *dicrawling* sebelumnya. Sebagai contoh digunakan 8 dokumen sebagai data latih.

Tabel 4.20 Dokumen data latih

No.	Komentar
1.	Kemendikbut cerdas mari sikapi kebijakan zonasi PPDB dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut
2.	Tujuannya baik, tp pelaksanaannya sangat rumit, spt daerah kami, sekolah2 banyak yg belum sesuai dgn 8 SNP, oleh sebab itu Kemendikbud dan Pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana

3.	Sistim zonasi tidak efektif...kasian rumahnya yg agak jauh...akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar
4.	InshaaAllah mutu seluruh sekolah di Indonesia bisa sama baik negeri maupun swasta . . itulah salah satu maksud dan tujuannya
5.	Seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi. Faktanya banyak sekolah yg sarana dan prasarananya sangat2 minim terutama ruang kelas yg tidak memadai, akhirnya PBM pun berjalan semrawut
6.	Zonasi demi keadilan sebenarnya tidak adil.
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah, sistem zonasi gak guna.
8.	Sistim zonasi mngecewakan bnyak phak...kpingin sekola yg lbih dekat ja g bsa msuk zonasi

b. Memberikan sentimen secara manual setiap dokumennya

Tabel 4.21 Penentuan sentimen data latih

No.	Komentar	Sentimen
1.	Kemendikbut cerdas mari sikapi kebijakan zonasi PPDB dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut	Positif
2.	Tujuannya baik, tp pelaksanaannya sangat rumit, spt daerah kami, sekolah2 banyak yg belum sesuai dgn 8 SNP, oleh sebab itu Kemendikbud dan Pemerintah daerah hrs	Negatif

	berani membuat terobosan jgn sekedar wacana	
3.	Sistim zonasi tidak efektif...kasion rumahnya yg agak jauh...akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar	Negatif
4.	InshaaAllah mutu seluruh sekolah di Indonesia bisa sama baik negeri maupun swasta . . itulah salah satu maksud dan tujuannya	Positif
5.	Seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi. Faktanya banyak sekolah yg sarana dan prasarananya sangat2 minim terutama ruang kelas yg tidak memadai, akhirnya PBM pun berjalan semrawut	Negatif
6.	Zonasi demi keadilan sebenarnya tidak adil.	Negatif
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah, sistem zonasi gak guna.	Negatif
8.	Sistim zonasi mngecewakan banyak pihak...kpingin sekola yg lebih dekat ja g bsa masuk zonasi	Negatif

c. Melakukan proses *case folding*

Tabel 4.22 Hasil Proses *Case Folding*

No.	Komentar
1.	kemendikbut cerdas mari sikapi kebijakan zonasi ppdb dengan bijak dan pemerintah dlm hal ini kemendikbud juga

	harus siap mengatasi kendala mengenai sistem zonasi tersebut
2.	tujuannya baik, tp pelaksanaannya sangat rumit, spt daerah kami, sekolah2 banyak yg belum sesuai dgn 8 snp, oleh sebab itu kemendikbud dan pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana
3.	sistim zonasi tidak efektif...kasian rumahnya yg agak jauh...akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar
4.	inshaaallah mutu seluruh sekolah di indonesia bisa sama baik negeri maupun swasta . . itulah salah satu maksud dan tujuannya
5.	seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi. faktanya banyak sekolah yg sarana dan prasarananya sangat2 minim terutama ruang kelas yg tidak memadai, akhirnya pbm pun berjalan semrawut
6.	zonasi demi keadilan sebenarnya tidak adil.
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah, sistem zonasi gak guna.
8.	sistim zonasi mngecewakan bnyak phak...kpingin sekola yg lebih dekat ja g bsa masuk zonasi

d. Melakukan proses *filtering*

Tabel 4.23 Hasil Proses *Filtering*

No.	Komentar
1.	kemendikbut cerdas mari sikapi kebijakan zonasi ppdb dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut

2.	tujuannya baik tp pelaksanaannya sangat rumit spt daerah kami sekolah banyak yg belum sesuai dgn smp oleh sebab itu kemendikbud dan pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana
3.	sistim zonasi tidak efektif kasian rumahnya yg agak jauh akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yg serba bayar
4.	inshaaallah mutu seluruh sekolah di indonesia bisa sama baik negeri maupun swasta itulah salah satu maksud dan tujuannya
5.	seharusnya mendikbut kaji dulu dilapangan sebelum menerapkan sisten zonasi faktanya banyak sekolah yg sarana dan prasarannya sangat minim terutama ruang kelas yg tidak memadai akhirnya pbm pun berjalan semrawut
6.	zonasi demi keadilan sebenarnya tidak adil
7.	selama sarana prasarana dan sdm masih belum merata di tiap sekolah sistem zonasi gak guna
8.	sistim zonasi mngecewakan banyak pihak kpingin sekola yg lebih dekat ja g bsa masuk zonasi

e. Melakukan proses *tokenization*

Tabel 4.24 Hasil Proses *Tokenization*

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
kemendikbut	tujuannya	sistim	inshaaallah
cerdas	baik	zonasi	mutu
mari	tp	tidak	seluruh
sikapi	pelaksanaannya	efektif	sekolah
kebijakan	sangat	kasian	di
zonasi	rumit	rumahnya	indonesia
ppdb	spt	yg	bisa

dengan bijak dan pemerintah dlm hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut	daerah kami sekolah banyak yg belum sesuai dgn snp oleh sebab itu kemendikbud dan pemerintah daerah hrs berani membuat terobosan jgn sekedar wacana	agak jauh akhirnya mereka dengan berat hati masukin anakanya ke sekolah swasta yg serba bayar	sama baik negeri maupun swasta itulah salah satu maksud dan tujuannya
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
seharusnya mendikbut kaji dulu dilapangan sebelum	zonasi demi keadilan sebenarnya tidak adil	selama sarana prasarana dan sdm masih	sistim zonasi mncecewakan banyak pihak kpingin

menerapkan sisten zonasi faktanya banyak sekolah yg sarana dan prasarananya sangat minim terutama ruang kelas yg tidak memadai akhirnya pbm pun berjalan semrawut		belum merata di tiap sekolah sistem zonasi gak guna	sekola yg lebih dekat ja g bisa masuk zonasi
---	--	---	--

f. Melakukan proses normalisasi kata

Tabel 4.25 Hasil Proses Normalisasi Kata

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
kemendikbud cerdas mari sikapi	tujuannya baik tapi pelaksanaannya	sistem zonasi tidak efektif	insya allah mutu seluruh sekolah

kebijakan zonasi ppdb dengan bijak dan pemerintah dalam hal ini kemendikbud juga harus siap mengatasi kendala mengenai sistem zonasi tersebut	sangat rumit seperti daerah kami sekolah banyak yang belum sesuai dengan snp oleh sebab itu kemendikbud dan pemerintah daerah harus berani membuat terobosan jangan sekedar wacana	kasihan rumahnya yang agak jauh akhirnya mereka dengan berat hati masukin anaknya ke sekolah swasta yang serba bayar	di indonesia bisa sama baik negeri maupun swasta itulah salah satu maksud dan tujuannya
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
seharusnya kemendikbud kaji	zonasi demi keadilan	selama sarana prasarana	sistem zonasi mengecewakan

dahulu dilapangan sebelum menerapkan sistem zonasi faktanya banyak sekolah yang sarana dan prasarananya sangat minim terutama ruang kelas yang tidak memadai akhirnya pbm pun berjalan semrawut	sebenarnya tidak adil	dan sdm masih belum merata di tiap sekolah sistem zonasi tidak guna	banyak pihak kepingin sekolah yang lebih dekat saja tidak bisa masuk zonasi
--	-----------------------------	--	--

g. Melakukan proses *stemming*

Tabel 4.26 Hasil Proses *Stemming*

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
kemendikbud	tuju	sistem	insya allah

cerdas mari sikap bijak zonasi ppdb dengan bijak dan pemerintah dalam hal ini kemendikbud juga harus siap atas kendala kena sistem zonasi sebut	baik tapi laksana sangat rumit seperti daerah kami sekolah banyak yang belum sesuai dengan snp oleh sebab itu kemendikbud dan pemerintah daerah harus berani buat terobos jangan sekedar wacana	zonasi tidak efektif kasihan rumah yang agak jauh akhir mereka dengan berat hati masuk anak ke sekolah swasta yang serba bayar	mutu seluruh sekolah di indonesia bisa sama baik negeri maupun swasta itu salah satu maksud dan tuju
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)

harus	zonasi	selama	sistem
kemendikbud	demi	sarana	zonasi
kaji	adil	prasarana	kecewa
dahulu	benar	dan	banyak
lapang	tidak	sdm	pihak
belum	adil	masih	kepingin
terap		belum	sekolah
sistem		rata	yang
zonasi		di	lebih
fakta		tiap	dekat
banyak		sekolah	saja
sekolah		sistem	tidak
yang		zonasi	bisa
sarana		tidak	masuk
dan		guna	zonasi
prasarana			
sangat			
minim			
utama			
ruang			
kelas			
yang			
tidak			
ada			
akhir			
pbm			
pun			
jalan			
semrawut			

h. Melakukan proses *filtering* / *stopword*

Tabel 4.27 Hasil Proses *Filtering / Stopword*

Dokumen (1)	Dokumen (2)	Dokumen (3)	Dokumen (4)
cerdas	tuju	sistem	insya allah
mari	baik	zonasi	mutu
sikap	laksana	tidak	seluruh
bijak	rumit	efektif	sekolah
zonasi	daerah	kasihan	indonesia
bijak	sekolah	rumah	bisa
pemerintah	banyak	jauh	sama
harus	belum	akhir	baik
siap	sesuai	berat	negeri
atas	sebab	hati	swasta
kendala	pemerintah	masuk	salah
kena	daerah	anak	maksud
sistem	harus	sekolah	tuju
zonasi	berani	swasta	
sebut	buat	serba	
	terobos	bayar	
	sekedar		
	wacana		
Dokumen (5)	Dokumen (6)	Dokumen (7)	Dokumen (8)
harus	zonasi	selama	sistem
kaji	demi	sarana	zonasi
dahulu	adil	prasarana	kecewa
lapang	benar	sdm	banyak
belum	tidak	masih	pihak
terap	adil	belum	kepingin
sistem		rata	sekolah
zonasi		tiap	lebih
fakta		sekolah	dekat

banyak sekolah sarana prasarana minim utama ruang kelas tidak ada akhir jalan semrawut		sistem zonasi tidak guna	tidak bisa masuk zonasi
--	--	-----------------------------------	----------------------------------

- i. Dari proses *pre-processing* pada data latih, hasilnya dibuat *index* untuk disimpan oleh *variable*.
2. Menyiapkan data yang ingin diuji. Sebagai contoh akan dilakukan 1 data uji agar prosesnya terlihat.

Tidaak efektif karena belum semuanya sekolah di Indonesia bagus

3. Melakukan *pre-processing* pada data uji. Pada tahap awal akan melakukan proses *case folding*.

tidaak efektif karena belum semuanya sekolah di indonesia bagus

4. Melakukan proses *tokenization*

tidaak
efektif
karena
belum
semuanya

sekolah
di
indonesia
bagus

5. Melakukan proses normalisasi kata ejaan dengan bantuan algoritma *levenshtein distance* untuk *typo*. Disini terdapat 2 kata yang ingin diperbaiki, yaitu kata “tidaak” dan kata “efejtif”. Perbaiki kata yang ditargetkan yaitu kata “tidak” dan kata “efektif”. Penggunaan algoritma *Levenshtein Distance* ini menghitung jarak terkecil dengan rumus yang dapat dilihat pada rumus 2.3, 2.4, 2.5, dan 2.6.

Tabel 4.28 Perhitungan Matriks *Levenshtein Distance* 1

		T	I	D	A	K
	0	1	2	3	4	5
T	1	0	1	2	3	4
I	2	1	0	1	2	3
D	3	2	1	0	1	2
A	4	3	2	1	0	1
A	5	4	3	2	1	1
K	6	5	4	3	2	1

Tabel 4.29 Perhitungan Matriks *Levenshtein Distance* 2

		E	F	E	K	T	I	F
	0	1	2	3	4	5	6	7
E	1	0	1	2	3	4	5	6
F	2	1	0	1	2	3	4	5
E	3	2	1	0	1	2	3	4
J	4	3	2	1	1	2	3	4
T	5	4	3	2	2	1	2	3
I	6	5	4	3	2	1	1	2
F	7	6	5	4	3	2	1	1

6. Melakukan proses *stemming*, merubah kata berimbuhan mejadi kata dasar sesuai dengan KBBI

tidak
efektif
karena
belum
semua
sekolah
di
indonesia
bagus

7. Melakukan proses *filtering / stopword*.

tidak
efektif
karena
belum
semua
sekolah
indonesia
bagus

8. Melakukan tahap terakhir di *pre-processing* pada penelitian ini yaitu *indexing* pada data uji.
9. Melakukan pencarian atau pencocokan *index* yang ada di dalam data latih dan *index* yang ada di data uji. Didapatkan *index* data uji yang sama dengan data latih yaitu data ke 2, 3, 4, 5, 6, 7, dan 8 pada data latih.
10. Melakukan proses pembobotan menggunakan algoritma *TF-IDF*. Proses perhitungan awal yaitu menghitung *IDF* dengan rumus yang dapat dilihat pada rumus 2.7.

Tabel 4.30 Hasil Perhitungan *IDF* Skenario 2

IDF	TF-0	TF-1	TF-2	TF-3	TF-4	TF-5	TF-6	TF-7	DF
-----	------	------	------	------	------	------	------	------	----

tidak	0.146128	1	0	1	0	1	1	1	1	5
efektif	0.845098	1	0	1	0	0	0	0	0	1
karena	0	1	0	0	0	0	0	0	0	0
belum	0.367977	1	1	0	0	1	0	0	1	3
semua	0	1	0	0	0	0	0	0	0	0
sekolah	0.066947	1	1	1	1	1	0	1	1	6
indonesia	0	1	0	0	0	0	0	0	0	0
baik	0.544068	1	1	0	1	0	0	0	0	2
tuju	0.544068	0	1	0	1	0	0	0	0	2
laksana	0.845098	0	1	0	0	0	0	0	0	1
rumit	0.845098	0	1	0	0	0	0	0	0	1
daerah	0.845098	0	2	0	0	0	0	0	0	1
banyak	0.367977	0	1	0	0	1	0	0	1	3
sesuai	0.845098	0	1	0	0	0	0	0	0	1
sebab	0.845098	0	1	0	0	0	0	0	0	1
pemerintah	0.845098	0	1	0	0	0	0	0	0	1
harus	0.544068	0	1	0	0	1	0	0	0	2
berani	0.845098	0	1	0	0	0	0	0	0	1
buat	0.845098	0	1	0	0	0	0	0	0	1
terobos	0.845098	0	1	0	0	0	0	0	0	1
sekedar	0.845098	0	1	0	0	0	0	0	0	1
wacana	0.845098	0	1	0	0	0	0	0	0	1
sistem	0.243038	0	0	1	0	1	0	1	1	4
zonasi	0.146128	0	0	1	0	1	1	1	1	5
kasihan	0.845098	0	0	1	0	0	0	0	0	1
rumah	0.845098	0	0	1	0	0	0	0	0	1
jauh	0.845098	0	0	1	0	0	0	0	0	1
akhir	0.544068	0	0	1	0	1	0	0	0	2
berat	0.845098	0	0	1	0	0	0	0	0	1
hati	0.845098	0	0	1	0	0	0	0	0	1
masuk	0.544068	0	0	1	0	0	0	0	1	2
anak	0.845098	0	0	1	0	0	0	0	0	1
swasta	0.544068	0	0	1	1	0	0	0	0	2
serba	0.845098	0	0	1	0	0	0	0	0	1
bayar	0.845098	0	0	1	0	0	0	0	0	1
insya allah	0.845098	0	0	0	1	0	0	0	0	1
mutu	0.845098	0	0	0	1	0	0	0	0	1
seluruh	0.845098	0	0	0	1	0	0	0	0	1
bisa	0.544068	0	0	0	1	0	0	0	1	2
sama	0.845098	0	0	0	1	0	0	0	0	1
negeri	0.845098	0	0	0	1	0	0	0	0	1

salah	0.845098	0	0	0	1	0	0	0	0	1
maksud	0.845098	0	0	0	1	0	0	0	0	1
kaji	0.845098	0	0	0	0	1	0	0	0	1
dahulu	0.845098	0	0	0	0	1	0	0	0	1
lapang	0.845098	0	0	0	0	1	0	0	0	1
terap	0.845098	0	0	0	0	1	0	0	0	1
fakta	0.845098	0	0	0	0	1	0	0	0	1
sarana	0.544068	0	0	0	0	1	0	1	0	2
prasarana	0.544068	0	0	0	0	1	0	1	0	2
minim	0.845098	0	0	0	0	1	0	0	0	1
utama	0.845098	0	0	0	0	1	0	0	0	1
ruang	0.845098	0	0	0	0	1	0	0	0	1
kelas	0.845098	0	0	0	0	1	0	0	0	1
ada	0.845098	0	0	0	0	1	0	0	0	1
jalan	0.845098	0	0	0	0	1	0	0	0	1
semrawut	0.845098	0	0	0	0	1	0	0	0	1
selama	0.845098	0	0	0	0	0	0	1	0	1
sdm	0.845098	0	0	0	0	0	0	1	0	1
masih	0.845098	0	0	0	0	0	0	1	0	1
rata	0.845098	0	0	0	0	0	0	1	0	1
tiap	0.845098	0	0	0	0	0	0	1	0	1
guna	0.845098	0	0	0	0	0	0	1	0	1
kecewa	0.845098	0	0	0	0	0	0	1	0	1
pihak	0.845098	0	0	0	0	0	0	1	0	1
kepingin	0.845098	0	0	0	0	0	0	1	0	1
lebih	0.845098	0	0	0	0	0	0	1	0	1
dekat	0.845098	0	0	0	0	0	0	1	0	1
adil	0.845098	0	0	0	0	0	2	0	0	1
benar	0.845098	0	0	0	0	0	1	0	0	1
demi	0.845098	0	0	0	0	0	1	0	0	1

Setelah mendapatkan nilai *IDF*, akan menghitung nilai bobotnya (*term weighting*) dengan rumus yang dapat dilihat pada rumus 2.8.

Tabel 4.31 Hasil Perhitungan bobot *TF-IDF* Skenario 2

	W-0	W-1	W-2	W-3	W-4	W-5	W-6	W-7
tidak	0.146128	0	0.146128	0	0.146128	0.146128	0.146128	0.146128
efektif	0.845098	0	0.845098	0	0	0	0	0
karena	0	0	0	0	0	0	0	0
belum	0.367977	0.367977	0	0	0.367977	0	0	0.367977
semua	0	0	0	0	0	0	0	0

sekolah	0.066947	0.066947	0.066947	0.066947	0.066947	0	0.066947	0.066947
indonesia	0	0	0	0	0	0	0	0
baik	0.544068	0.544068	0	0.544068	0	0	0	0
tuju	0	0.544068	0	0.544068	0	0	0	0
laksana	0	0.845098	0	0	0	0	0	0
rumit	0	0.845098	0	0	0	0	0	0
daerah	0	1.690196	0	0	0	0	0	0
banyak	0	0.367977	0	0	0.367977	0	0	0.367977
sesuai	0	0.845098	0	0	0	0	0	0
sebab	0	0.845098	0	0	0	0	0	0
pemerintah	0	0.845098	0	0	0	0	0	0
harus	0	0.544068	0	0	0.544068	0	0	0
berani	0	0.845098	0	0	0	0	0	0
buat	0	0.845098	0	0	0	0	0	0
terobos	0	0.845098	0	0	0	0	0	0
sekedar	0	0.845098	0	0	0	0	0	0
wacana	0	0.845098	0	0	0	0	0	0
sistem	0	0	0.243038	0	0.243038	0	0.243038	0.243038
zonasi	0	0	0.146128	0	0.146128	0.146128	0.146128	0.146128
kasihan	0	0	0.845098	0	0	0	0	0
rumah	0	0	0.845098	0	0	0	0	0
jauh	0	0	0.845098	0	0	0	0	0
akhir	0	0	0.544068	0	0.544068	0	0	0
berat	0	0	0.845098	0	0	0	0	0
hati	0	0	0.845098	0	0	0	0	0
masuk	0	0	0.544068	0	0	0	0	0.544068
anak	0	0	0.845098	0	0	0	0	0
swasta	0	0	0.544068	0.544068	0	0	0	0
serba	0	0	0.845098	0	0	0	0	0
bayar	0	0	0.845098	0	0	0	0	0
insya allah	0	0	0	0.845098	0	0	0	0
mutu	0	0	0	0.845098	0	0	0	0
seluruh	0	0	0	0.845098	0	0	0	0
bisa	0	0	0	0.544068	0	0	0	0.544068
sama	0	0	0	0.845098	0	0	0	0
negeri	0	0	0	0.845098	0	0	0	0
salah	0	0	0	0.845098	0	0	0	0
maksud	0	0	0	0.845098	0	0	0	0
kaji	0	0	0	0	0.845098	0	0	0
dahulu	0	0	0	0	0.845098	0	0	0
lapang	0	0	0	0	0.845098	0	0	0

terap	0	0	0	0	0.845098	0	0	0
fakta	0	0	0	0	0.845098	0	0	0
sarana	0	0	0	0	0.544068	0	0.544068	0
prasarana	0	0	0	0	0.544068	0	0.544068	0
minim	0	0	0	0	0.845098	0	0	0
utama	0	0	0	0	0.845098	0	0	0
ruang	0	0	0	0	0.845098	0	0	0
kelas	0	0	0	0	0.845098	0	0	0
ada	0	0	0	0	0.845098	0	0	0
jalan	0	0	0	0	0.845098	0	0	0
semrawut	0	0	0	0	0.845098	0	0	0
selama	0	0	0	0	0	0	0.845098	0
sdm	0	0	0	0	0	0	0.845098	0
masih	0	0	0	0	0	0	0.845098	0
rata	0	0	0	0	0	0	0.845098	0
tiap	0	0	0	0	0	0	0.845098	0
guna	0	0	0	0	0	0	0.845098	0
kecewa	0	0	0	0	0	0	0.845098	0
pihak	0	0	0	0	0	0	0.845098	0
kepingin	0	0	0	0	0	0	0.845098	0
lebih	0	0	0	0	0	0	0.845098	0
dekat	0	0	0	0	0	0	0.845098	0
adil	0	0	0	0	0	1.690196	0	0
benar	0	0	0	0	0	0.845098	0	0
demi	0	0	0	0	0	0.845098	0	0
TOTAL	1.970218	12.57628	9.840327	8.158905	13.65564	3.672648	10.98646	2.426331

Dimana total *term weighing* yang didapat pada perhitungan diatas yaitu:

Data uji = 1.970218

Kalimat – 1 = 12.57628

Kalimat – 2 = 9.840327

Kalimat – 3 = 8.158905

Kalimat – 4 = 13.65564

Kalimat – 5 = 3.672648

Kalimat – 6 = 10.98646

Kalimat – 7 = 2.426331

11. Melakukan proses *clustering* menggunakan algoritma *k-means* dimana prosesnya dapat dilihat pada gambar 4.9 dan penjelasannya bisa dilihat pada sub-bab 2.8.2. Data yang di *cluster* yaitu jumlah *Term. Weighting* dari setiap data yang sudah dihitung pada proses pembobotan.

- a. Menentukan jumlah *cluster*. Dipenelitian ini menggunakan jumlah *cluster* = 2. Karena *output* yang dihasilkan hanya sentimen positif dan negatif.

cluster 1

cluster 2

--	--

- b. Menentukan centroid awal secara random. Dimana dipenelitian ini centroid awal dari setiap *cluster* yaitu menggunakan nilai terkecil dan nilai terbesar dari jumlah total *term weighting*. Centroid pada *cluster 1* yaitu 1.970218 (data uji) dan centroid pada *cluster 2* yaitu 13.65564 (kalimat – 4)

cluster 1

cluster 2

1.970218 (c.a)	13.65564 (c.a)
----------------	----------------

Keterangan :

c.a = centroid awal

- c. Menghitung jarak objek ke centroid menggunakan *euclidean distance* untuk mendapatkan jarak terdekat (rumus bisa dilihat pada rumus 2.1).

Tabel 4.32 Perhitungan *Euclidean Distance cluster 1* Skenario 2

	<i>Cluster 1</i>
Data uji	$\sqrt{(1.970218 - 1.970218)^2} = 0$
Kalimat – 1	$\sqrt{(12.57628 - 1.970218)^2} = 10.6060$
Kalimat – 2	$\sqrt{(9.840327 - 1.970218)^2} = 7.8701$
Kalimat – 3	$\sqrt{(8.158905 - 1.970218)^2} = 6.1886$
Kalimat – 4	$\sqrt{(13.65564 - 1.970218)^2} = 11.6854$
Kalimat – 5	$\sqrt{(3.672648 - 1.970218)^2} = 1.7024$
Kalimat – 6	$\sqrt{(10.98646 - 1.970218)^2} = 9.0162$
Kalimat – 7	$\sqrt{(2.426331 - 1.970218)^2} = 0.4561$

Tabel 4.33 Perhitungan *Euclidean Distance cluster 2* Skenario 2

	<i>Cluster 2</i>
Data uji	$\sqrt{(1.970218 - 13.65564)^2} = 11.6854$
Kalimat – 1	$\sqrt{(12.57628 - 13.65564)^2} = 1.0793$
Kalimat – 2	$\sqrt{(9.840327 - 13.65564)^2} = 3.8147$
Kalimat – 3	$\sqrt{(8.158905 - 13.65564)^2} = 5.4967$
Kalimat – 4	$\sqrt{(13.65564 - 13.65564)^2} = 0$

Kalimat – 5	$\sqrt{(3.672648 - 13.65564)^2} = 9.9829$
Kalimat – 6	$\sqrt{(10.98646 - 13.65564)^2} = 2.6691$
Kalimat – 7	$\sqrt{(2.426331 - 13.65564)^2} = 11.2293$

- d. Membagi semua objek ke *cluster*, jika jarak objek dengan centroid pada *cluster 1* lebih kecil dibanding jarak objek dengan centroid pada *cluster 2* maka objek tersebut masuk ke dalam *cluster 1*, dan sebaliknya.

<i>cluster 1</i>	<i>cluster 2</i>
1.970218 (c.a)	13.65564 (c.a)
1.970218 (data uji)	12.57628 (kal-1)
3.672648 (kal-5)	9.840327 (kal-2)
2.426331 (kal-7)	8.158905 (kal-3)
	13.65564 (kal-4)
	10.98646 (kal-6)

Didapatkan pembagian *cluster* sesuai dengan perhitungan dari *euclidean distance* setiap objek sebelumnya yaitu seperti yang diatas.

- e. Mencari centroid baru dengan cara menghitung rata – rata nilai dari setiap *cluster*.

Centroid baru *cluster 1*:

$$(1.970218 + 3.672648 + 2.426331) / 3 = 2.6897$$

Centroid baru *cluster 2*:

$$(12.57628 + 9.840327 + 8.158905 + 13.65564 + 10.98646) / 5 = 8.8462$$

<i>cluster 1</i>	<i>cluster 2</i>
2.6897 (c.b)	8.8462 (c.b)

Keterangan :

c.b = centroid baru

- f. Menghitung kembali jarak antara objek dengan centroid baru dengan menggunakan *euclidean distance*.

Tabel 4.34 Perhitungan kembali *Euclidean Distance cluster 1*

	<i>Cluster 1</i>
Data uji	$\sqrt{(1.970218 - 2.6897)^2} =$ 0.7194
Kalimat – 1	$\sqrt{(12.57628 - 2.6897)^2} =$ 9.8865
Kalimat – 2	$\sqrt{(9.840327 - 2.6897)^2} =$ 7.1506
Kalimat – 3	$\sqrt{(8.158905 - 2.6897)^2} =$ 5.4692
Kalimat – 4	$\sqrt{(13.65564 - 2.6897)^2} =$ 10.9659
Kalimat – 5	$\sqrt{(3.672648 - 2.6897)^2} =$ 0.9829
Kalimat – 6	$\sqrt{(10.98646 - 2.6897)^2} =$ 9.2967
Kalimat – 7	$\sqrt{(2.426331 - 2.6897)^2} =$ 0.2633

Tabel 4.35 Perhitungan kembali *Euclidean Distance cluster 2*

	<i>Cluster 2</i>
Data uji	$\sqrt{(1.970218 - 8.8462)^2} = 6.8759$
Kalimat – 1	$\sqrt{(12.57628 - 8.8462)^2} = 3.7300$
Kalimat – 2	$\sqrt{(9.840327 - 8.8462)^2} = 0.9941$
Kalimat – 3	$\sqrt{(8.158905 - 8.8462)^2} = 0.6872$
Kalimat – 4	$\sqrt{(13.65564 - 8.8462)^2} = 4.8094$
Kalimat – 5	$\sqrt{(3.672648 - 8.8462)^2} = 5.1735$
Kalimat – 6	$\sqrt{(10.98646 - 8.8462)^2} = 2.1402$
Kalimat – 7	$\sqrt{(2.426331 - 8.8462)^2} = 6.4198$

- i. Membagi semua objek ke *cluster*, sama seperti sebelumnya jika jarak objek dengan centroid pada *cluster 1* lebih kecil dibanding jarak objek dengan centroid pada *cluster 2* maka objek tersebut masuk ke dalam *cluster 1*, dan sebaliknya. Disini akan dilihat apakah hasil pembagiannya terjadi perubahan atau tidak dari pembagian sebelumnya.

cluster 1

cluster 2

2.6897 (c.b)	8.8462 (c.b)
1.970218 (data uji)	12.57628 (kal-1)
3.672648 (kal-5)	9.840327 (kal-2)

2.426331 (kal-7)	8.158905 (kal-3)
	13.65564 (kal-4)
	10.98646 (kal-6)

Dari hasil pembagian dengan menghitung jarak antara objek dengan centroid baru, tidak terjadi perubahan *cluster* atau objeknya tidak ada yang berubah dari pembagian yang awal.

- j. Karena tidak terjadi perubahan atau perpindahan *cluster* dari setiap objeknya, maka iterasi dari proses *K-Means* berhenti. Selesai.

12. Mencari di *cluster* mana posisi objek dari data uji berada. Apakah di *cluster* 1 atau di *cluster* 2. Jika posisi objek dari data uji berada di *cluster* 1, maka *cluster* 2 dibiarkan atau tidak digunakan. Dan sebaliknya.

<i>cluster 1</i>	<i>cluster 2</i>
2.6897 (c.b)	8.8462 (c.b)
1.970218 (data uji)	12.57628 (kal-1)
3.672648 (kal-5)	9.840327 (kal-2)
2.426331 (kal-7)	8.158905 (kal-3)
	13.65564 (kal-4)
	10.98646 (kal-6)

Ditemukan objek dari data uji berada di *cluster* 1, maka *cluster* 2 tidak digunakan dan akan fokus pada *cluster* 1 saja.

13. Mencari sentimennya yaitu dengan cara dilihat apakah objek kalimat – n pada *cluster* tersebut lebih banyak bersentimen positif atau negatif. Jika lebih banyak positif, maka sentimen dari data uji yaitu positif, dan sebaliknya. Namun jika jumlah sentimen positif dan negatifnya sama, maka hasil sentimennya yaitu netral.

cluster 1

2.6897 (c.b)
1.970218 (data uji)
3.672648 (kal-5)
2.426331 (kal-7)

Dilihat dari anggota atau objek yang berada di *cluster 1* ada 3 yaitu data uji (yang ingin dicari sentimennya), objek kalimat – 5 yang bersentimen negatif, dan objek kalimat ke – 7 yang bersentimen negatif. Maka data uji yang dicari sentimennya juga bersentimen negatif.

14. Proses selanjutnya yaitu pelabelan secara manual pada data uji. Dilihat dari susunan katanya, data uji ini diberi sentimen negatif.

15. Menghitung akurasi antara hasil sentimen menggunakan algoritma *K-Means* dan hasil sentimen secara manual menggunakan *confusion matrix*. Rumusnya bisa dilihat pada rumus 2.9.

	Positive (+)	Negative (-)
Positive (+)	0	0
Negative (-)	0	1

$$\frac{\text{Jumlah data yang diprediksi benar}}{\text{Total data yang diprediksi}} = \frac{(0+1)}{(0+0+1+0)} \times 100\% = 100\%$$

16. Selesai.

4.5 Simulasi (*Simulation*)

Pada tahapan simulasi ini akan dilakukan simulasi aplikasi yang berkaitan dengan hasil sentimen dan pengujian tingkat akurasi kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance*. Adapun faktor-faktor dalam proses simulasi dapat dilihat pada tabel 4.36 berikut.

Tabel 4.36 Faktor Simulasi Penelitian

Variabel / Parameter Simulasi	Tahap Simulasi
Faktor 1	Tahap <i>clustering</i> sentimen data latih menggunakan cara manual.
Faktor 2	Tahap pelatihan data pada data latih
Faktor 3	Tahap pengujian data uji dengan kombinasi algoritma <i>K-Means</i> dan algoritma <i>Levensthein Distance</i> berdasarkan nilai k pada algoritma <i>K-Means</i> adalah 2
Faktor 4	Tahap pengujian akurasi menggunakan model confusion matrix

Variabel atau parameter simulasi yang digunakan dalam melakukan scenario – skenario pengujian yaitu dengan menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance*. Algoritma *K-Means* digunakan untuk melakukan pengklusteran pada data yang ingin diuji. Sedangkan algoritma *Levensthein Distance* digunakan sebagai algoritma normalisasi kata seperti kata – kata yang typo.

Dalam sub-sub berikut akan dibahas simulasi aplikasi sentimen berdasarkan skenario pada tabel 4.36.

4.5.1 Tahap Pengujian Data Uji

Pengujian yang dilakukan adalah penggunaan algoritma *K-Means* yaitu pada skenario 1 dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* yaitu pada skenario 2 untuk menentukan sentimen positif dan negatif.

Pada tahapan ini terdapat informasi sejauh mana tingkat keberhasilan algoritma tersebut di dalam data uji ke dalam skenario yang dihitung berdasarkan tingkat akurasi.

1. Skenario 1 (Membandingkan Hasil Sentimen Algoritma *K-Means* dan Hasil Sentimen secara Manual)

Pada skenario 1 ini didapatkan dari hasil perbandingan sentimen algoritma *K-Means* dengan nilai $k = 2$ dan sentimen secara manual yaitu sebesar 84%. Hasil dari aplikasi bisa dilihat pada gambar 4.11

Hasil Akurasi Algoritma K-Means	
Jumlah Data Uji	: 50
Jumlah Kesalahan	: 8
Persentase Validasi	: 84,00%

Gambar 4.11 Hasil Akurasi dari Aplikasi Skenario 1

2. Skenario 2 (Membandingkan Hasil Sentimen Kombinasi Algoritma *K-Means* dan Algoritma *Levensthein Distance* dan Hasil Sentimen secara Manual)

Pada skenario 1 ini didapatkan dari hasil perbandingan sentimen kombinasi algoritma *K-Means* dengan nilai $k = 2$ dan Algoritma *Levensthein Distance* dan sentimen secara manual yaitu sebesar 90%. Hasil dari aplikasi bisa dilihat pada gambar 4.12.

Hasil Akurasi Algoritma K-Means dan Algoritma Levensthein Distance	
Jumlah Data Uji	: 50
Jumlah Kesalahan	: 5
Persentase Validasi	: 90,00%

Gambar 4.12 Hasil Akurasi dari Aplikasi Skenario 2

4.6 Verifikasi dan Validasi (*Verification and Validation*)

Pembahasan untuk sub-bab ini akan dibahas pada bab 5.

4.7 Eksperimentasi (*Experimentation*)

Pembahasan untuk sub-bab ini akan dibahas pada bab 5.

4.8 Analisis Keluaran (*Output Analysis*)

Pembahasan untuk sub-bab ini akan dibahas pada bab 5



BAB 5

HASIL DAN PEMBAHASAN

5.1 Verifikasi dan Validasi (*Verification and Validation*)

Verifikasi dilakukan untuk memastikan bahwa setiap tahapan pada bab-bab sebelumnya saling memiliki hubungan, dalam hal ini setiap tahapan pada bab 4 diulas kembali untuk memastikan tiap tahap tersebut saling terkait. Verifikasi juga memastikan bahwa input dan output sesuai dengan yang diharapkan dimulai dari tahap *problem formulation* (formulasi masalah) hingga *simulation phase* (simulasi).

Pada tahapan formulasi masalah (*problem formulation*) dilakukan pembahasan mengenai masalah terpenting dengan cara identifikasi masalah untuk dirumuskan dalam penulisan skripsi ini, sehingga dari permasalahan tersebut dapat dikembangkan suatu pemodelan konsep sebagai solusi. Selanjutnya pada tahapan model pengkosepan (*conceptual model*), dilakukan pembahasan konsep secara keseluruhan pada aplikasi analisis sentimen meliputi *input*, proses, eksperimen dan *output* yang diharapkan. Pada tahapan data masukan / keluaran (*input / output data*), membahas *input* dan *output* data dengan menyebutkan atribut-atribut data yang akan disimpan ke dalam *database* Mysql. Berlanjut ke tahapan pemodelan (*Modelling*) yaitu berkaitan dengan mengolah data *input* dan *output* yang telah dibuat pada tahapan sebelumnya. Pada tahapan ini dilakukan perhitungan sampel dan konstruksi sentimen data latih secara manual, perhitungan sampel dan konstruksi klasifikasi data uji menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* secara manual untuk dapat dijadikan acuan dalam pembuatan aplikasi pada skripsi ini. Berikutnya pada tahapan terakhir adalah simulasi (*simulation*), yaitu dengan melakukan simulasi pada aplikasi analisis sentimen yang fungsinya mengimplementasikan pemodelan-pemodelan manual sebelumnya. Oleh karena itu, setiap

tahapan dapat dipastikan memiliki keterkaitan, karena setiap tahapan yang dibuat akan berpengaruh untuk membuat tahapan selanjutnya. Sehingga dari seluruh tahapan-tahapan yang dibahas pada sub-bab sebelumnya dapat diverifikasi sesuai dengan ketentuan verifikasi yang ada.

Dalam proses validasi dilakukan pengujian kebenaran sistem yaitu dengan melakukan perbandingan hasil algoritma *K-Means* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* yang dihitung secara manual dengan hasil pada aplikasi analisis sentimen sehingga menghasilkan keakuratan sistem.

5.2 Eksperimentasi (*Experimentation*)

Eksperimen yang dilakukan yaitu dengan membandingkan hasil skenario yaitu hasil sentimen data uji menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance*. Parameter yang digunakan untuk nilai k pada algoritma *K-Means* yaitu $k=2$. Dari eksperimen tersebut dilakukan analisis outputnya yang akan dibahas pada tahapan *Output Analysis*.

5.3 Analisis Keluaran (*Output Analysis*)

Pada tahap analisis keluaran, dilakukan analisis terhadap hasil sentimen masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* beserta hasil tingkat akurasi dari setiap scenario tersebut. *Output* penelitian ini didapatkan dari aplikasi yang sudah dibangun menggunakan algoritma tersebut. Hasil sentimen dan tingkat akurasi dari setiap algoritma yang digunakan akan dijelaskan pada sub-bab 5.3.1 , sub-bab 5.3.2 , dan sub-bab 5.3.3.

5.3.1 Hasil Sentimen Algoritma *K-Means* dan Kombinasi Algoritma *K-Means* dan Algoritma *Levensthein Distance*

Pada sub-bab ini menjelaskan dari hasil sentimen 50 data uji yang bersumber dari YouTube dan Facebook yaitu komentar

masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah. Pengujian dilakukan menggunakan algoritma *K-Means* dan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* dibandingkan dengan hasil sentimen yang dilakukan secara manual sebanyak 50 data uji.

Tabel 5.1 Hasil Sentimen dari Skenario Pada Data Uji

Data ke-n	Algoritma <i>K-Means</i>	Kombinasi Algoritma <i>K-Means</i> dan Algoritma <i>Levensthein Distance</i>	Sentimen sebenarnya
1	Positif	Positif	Positif
2	Positif	Positif	Positif
3	Positif	Negatif	Positif
4	Negatif	Negatif	Positif
5	Positif	Positif	Positif
6	Negatif	Negatif	Positif
7	Negatif	Positif	Positif
8	Positif	Positif	Positif
9	Positif	Positif	Positif
10	Negatif	Negatif	Positif
11	Negatif	Positif	Positif
12	Negatif	Negatif	Negatif
13	Positif	Positif	Positif
14	Positif	Positif	Positif
15	Negatif	Positif	Positif
16	Negatif	Negatif	Negatif
17	Negatif	Negatif	Negatif
18	Negatif	Negatif	Negatif
19	Negatif	Negatif	Negatif
20	Negatif	Negatif	Negatif
21	Negatif	Negatif	Negatif

22	Negatif	Negatif	Negatif
23	Negatif	Negatif	Negatif
24	Negatif	Negatif	Negatif
25	Negatif	Negatif	Negatif
26	Negatif	Negatif	Negatif
27	Negatif	Negatif	Negatif
28	Positif	Negatif	Negatif
29	Negatif	Negatif	Negatif
30	Negatif	Negatif	Negatif
31	Negatif	Negatif	Negatif
32	Negatif	Negatif	Negatif
33	Negatif	Negatif	Negatif
34	Negatif	Positif	Negatif
35	Negatif	Negatif	Negatif
36	Negatif	Negatif	Negatif
37	Negatif	Negatif	Negatif
38	Negatif	Negatif	Negatif
39	Negatif	Negatif	Negatif
40	Positif	Negatif	Negatif
41	Negatif	Negatif	Negatif
42	Negatif	Negatif	Negatif
43	Negatif	Negatif	Negatif
44	Negatif	Negatif	Negatif
45	Negatif	Negatif	Negatif
46	Negatif	Negatif	Negatif
47	Negatif	Negatif	Negatif
48	Negatif	Negatif	Negatif
49	Negatif	Negatif	Negatif
50	Negatif	Negatif	Negatif

5.3.2 Analisis Hasil Akurasi Algoritma *K-Means*

Pada skenario 1 yaitu menganalisis sentimen masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah menggunakan algoritma *K-Means* dan menghitung tingkat akurasi dari penggunaan algoritma *K-Means*. Hasil pengujian dari skenario 1 ini dapat dilihat pada tabel 5.2.

Tabel 5.2 Hasil Pengujian Skenario 1

Actual Value Prediction Value	Positive (+)	Negative (-)
Positive (+)	True positive = 8	False positive = 2
Negative (-)	False negative = 6	True negative = 34

Berdasarkan hasil pengujian pada skenario 1 dari tabel diatas dapat diambil kesimpulan sebagai berikut:

1. Sentimen dari masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah yaitu lebih banyak yang bersentimen negatif dibanding yang bersentimen positif.
2. Nilai akurasi dari skenario 1 yaitu perbandingan antara hasil sentimen dari algoritma *K-Means* tanpa menggunakan algoritma normalisasi kata dan hasil sentimen secara manual mendapatkan tingkat akurasi yang lebih kecil dibandingkan skenario 2, perhitungan hasil bisa dilihat dibawah ini.

$$\frac{\text{Jumlah data yang diprediksi benar}}{\text{Total data yang diprediksi}} = \frac{(8 + 34)}{(8 + 2 + 34 + 6)} \times 100 = 84\%$$

5.3.3 Analisis Hasil Akurasi Kombinasi Kombinasi Algoritma *K-Means* dan Algoritma *Levenshtein Distance* dan Sentimen Manual

Pada skenario 2 yaitu menganalisis sentimen masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah menggunakan kombinasi algoritma *K-Means* dan algoritma *Levenshtein Distance* dan menghitung tingkat akurasi dari penggunaan kombinasi algoritma *K-Means* dan algoritma *Levenshtein Distance*. Hasil pengujian dari skenario 2 ini dapat dilihat pada tabel 5.3.

Tabel 5.3 Hasil Pengujian Skenario 2

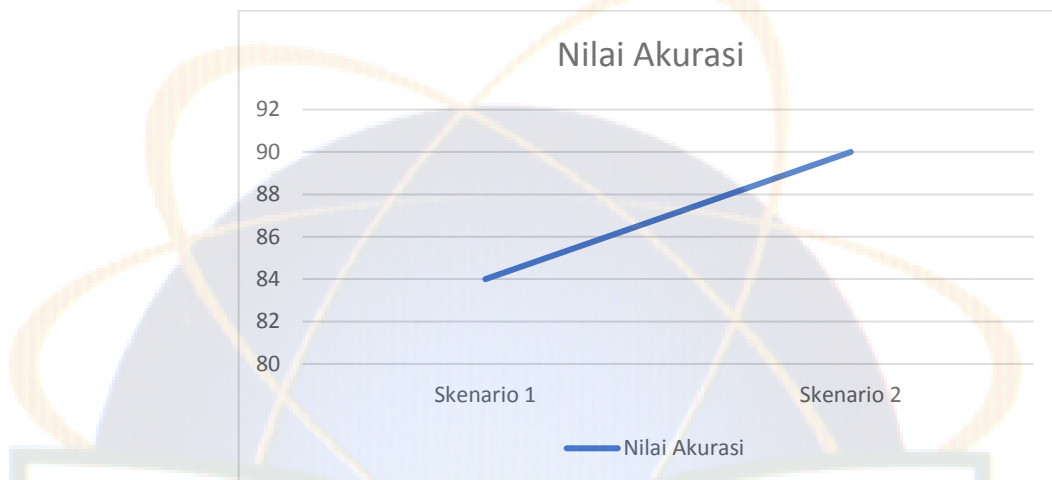
Actual Value Prediction Value	Positive (+)	Negative (-)
Positive (+)	True positive = 10	False positive = 1
Negative (-)	False negative = 4	True negative = 35

Berdasarkan hasil pengujian pada skenario 2 dari tabel diatas dapat diambil kesimpulan sebagai berikut:

1. Sentimen dari masyarakat terhadap kebijakan pemerintah tentang sistem zonasi sekolah yaitu lebih banyak yang bersentimen negatif dibanding yang bersentimen positif.
2. Nilai akurasi dari skenario 2 yaitu perbandingan antara hasil sentimen dari kombinasi algoritma *K-Means* dan algoritma normalisasi kata yaitu algoritma *Levenshtein distance* dan hasil sentimen secara manual mendapatkan tingkat akurasi lebih tinggi dibandingkan skenario 1 karena menggunakan algoritma normalisasi kata ini. Perhitungan hasil bisa dilihat dibawah ini.

$$\frac{\text{Jumlah data yang diprediksi benar}}{\text{Total data yang diprediksi}} = \frac{(10 + 35)}{(10 + 1 + 35 + 4)} \times 100 = 90\%$$

5.3.4 Analisis Hasil Peningkatan Akurasi Kombinasi Algoritma *K-Means* dan Algoritma *Levensthein Distance* dibandingkan dengan Algoritma *K-Means* saja.



Gambar 5.1 Grafik Peningkatan Akurasi Pada Skenario 2

Berdasarkan gambar grafik 5.1 maka didapatkan hasil analisis pada kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma *K-Means* saja. Pada skenario 1 (menggunakan algoritma *K-Means* saja) mendapatkan nilai akurasi sebesar 84%. Pada skenario 2 (menggunakan algoritma *K-Means* dan algoritma *Levensthein Distance*) mendapatkan nilai akurasi sebesar 90%. Hal ini disebabkan karena di dalam data uji ditemukan suatu kata yang typo dan di skenario 1 ini tidak adanya proses normalisasi kata, setelah dilakukan pencocokan data pada data latih yang mengandung kata tersebut (*term frekuensi*) = 0. Berbeda halnya pada skenario 2, kata yang typo ini akan dilakukan normalisasi kata menggunakan algoritma *Levensthein Distance*. Hasilnya untuk kata yang telah diperbaiki oleh algoritma *Levensthein Distance* ini setelah dilakukan pencocokan data pada

data latih yang mengandung kata tersebut (*term frekuensi*) > 0 . Hasil pencocokan itu berpengaruh pada nilai pembobotan kata dan inilah yang menyebabkan hasil sentimen dari proses *K-Means* mengalami perubahan dan peningkatan akurasi dibandingkan dengan skenario 1



BAB 6

PENUTUP

6.1 Kesimpulan

Berdasarkan masalah yang telah dirumuskan di bab 1 pada sub-bab 1.2 dan telah dideskripsikan pada bab-bab sebelumnya, peneliti mendapatkan hasil yaitu tingkat akurasi yang didapatkan dari 2 skenario yaitu pada skenario 1 peneliti hanya menggunakan algoritma *K-Means* tanpa bantuan algoritma normalisasi kata didapatkan nilai akurasinya yaitu 84% dan pada skenario 2 peneliti menggunakan kombinasi algoritma *K-Means* dan algoritma *Levensthein Distance* sebagai algoritma normalisasi kata mengalami peningkatan akurasi yaitu 90%.

Kesimpulan yang bisa diambil dari hasil di atas yaitu pada penelitian analisis teks diharuskan menggunakan fitur tambahan berupa normalisasi kata yang baik. Masyarakat pada umumnya, mereka menggambarkan ekspresi ataupun pendapat mereka menggunakan kata – kata singkatan yang tidak ada di dalam KBBI (jika berbahasa Indonesia). Dengan menggunakan fitur normalisasi kata ini, peneliti bisa mengekstrak informasi dengan lebih jelas dan akurat dari data – data yang sudah diambil melalui target sumber.

6.2 Saran

Pada penelitian saat ini peneliti menyadari bahwa masih banyak kekurangan dan keterbatasan. Oleh karena itu, ada beberapa hal yang bisa disarankan untuk penelitian selanjutnya agar hasilnya lebih memuaskan dan lebih baik, yaitu:

1. Sistem ini hanya bisa menggunakan bahasa Indonesia. Diharapkan penelitian selanjutnya bisa menggunakan selain bahasa Indonesia.
2. Menambahkan data yang akan dianalisis agar hasilnya lebih meyakinkan.

3. Menggunakan *k-fold cross validation* untuk menghitung nilai akurasi.



DAFTAR PUSTAKA

- A.S, R., & Shalahuddin, M. (2014). *Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek* (2nd ed.). Bandung: Informatika Bandung.
- Antinasari, P., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku, *1*, 1733–1741. Retrieved from https://www.researchgate.net/publication/320234777_Analisis_Sentimen_Tentang_Opini_Film_pada_Dokumen_Twitter_Berbahasa_Indonesia_Menggunakan_Naive_Bayes_dengan_Perbaikan_Kata_Tidak_Baku
- Bintana, R. R., & Agustian, S. (2012). Penerapan Model OKAPI BM25 Pada Sistem Temu Kembali Informasi. Retrieved from <http://ejournal.uin-suska.ac.id/index.php/SNTIKI/article/view/2905>
- Budi, S. (2017). Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algoritma K-Means, *16*. Retrieved from <https://publikasi.dinus.ac.id/index.php/technoc/article/viewFile/1263/1025%0A>
- Faisal, M. R. (2017). *Seri Belajar Data Science: Supervised Larning dengan R* (1st ed.). Banjarmasin: PE Press.
- Kemendikbud, P. W. (2018). Kemendikbud: Sistem Zonasi Sangat Tepat untuk Pemerataan Pendidikan. Retrieved from <https://www.kemdikbud.go.id/main/blog/2018/07/kemdikbud-sistem-zonasi-sangat-tepat-untuk-pemerataan-pendidikan>
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool. Retrieved from <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
- Madani, S., Kazmi, J., & Mahlknecht, S. (2010). Wireless sensor networks: modeling and simulation. Retrieved from <https://pdfs.semanticscholar.org/2a16/af9d99abe589a8979c1fbc17ad7c237c1f3b.pdf>
- Merliana, N. P. E., Ernawati, & Santoso, A. (2015). ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING. Retrieved from <https://media.neliti.com/media/publications/174505-ID-none.pdf>
- Nugroho, G. A. P. (2016). *Analisis Sentimen Data Twitter Menggunakan K-Means Clustering*. UNIVERSITAS SANATA DHARMA YOGYAKARTA.
- Okfalisa, & Harahap, A. H. (2016). Implementasi Metode Terms Frequency-Inverse Document Frequency (TF-IDF) dan Maximum Marginal Relevance untuk Monitoring Diskusi Online, *13*. Retrieved from <http://ejournal.uin->

suska.ac.id/index.php/sitekin/article/viewFile/1399/1408

- Prasidhatama, A., & Suryaningrum, K. M. (2018). Perbandingan Algoritma Nazief & Adriani Dengan Algoritma Idris Untuk Pencarian Kata Dasar, 4. Retrieved from <http://jurnal.unmer.ac.id/index.php/jtmi/article/view/1773/1157>
- Pratama, B. P., & Pamungkas, S. A. (2016). ANALISIS KINERJA ALGORITMA LEVENSHTAIN DISTANCE DALAM MENDETEKSI KEMIRIPAN DOKUMEN TEKS, 6 No.2. Retrieved from <http://journal.uinjkt.ac.id/index.php/logika/article/view/3917>
- Rohmawati, N., Defiyanti, S., & Jajuli, M. (2015). IMPLEMENTASI ALGORITMA K-MEANS DALAM PENGKLASTERAN MAHASISWA PELAMAR BEASISWA. Retrieved from https://www.researchgate.net/publication/294260664_IMPLEMENTASI_ALGORITMA_K-MEANS_DALAM_PENGKLASTERAN_MAHASISWA_PELAMAR_BEASISWA
- Rosandhy, C. (2017). *Sistem Analisis Sentimen Pada Komentar Evaluasi Dosen di SION STIKOM Bali Menggunakan Gabungan Metode K-Means dan K-Nearest Neighbor*. SION STIKOM Bali.
- Rosandy, T. (2016). PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus: KSPPS / BMT AL-FADHILA), 2. Retrieved from <https://media.neliti.com/media/publications/141765-ID-none.pdf>
- Salim, M. A., & Anistyasari, Y. (2017). PENGEMBANGAN APLIKASI PENILAIAN UJIAN ESSAY BERBASIS ONLINE MENGGUNAKAN ALGORITMA NAZIEF DAN ADRIANI DENGAN METODE COSINE SIMILARITY, 2. Retrieved from <http://jurnalmahasiswa.unesa.ac.id/index.php/it-edu/article/view/21338/19570>
- Saputro, D. (2016). *ANALISIS PERBANDINGAN ALGORITMA ROUTING LEACH-C DAN PEGASIS PADA WIRELESS SENSOR NETWORK MENGGUNAKAN NETWORK SIMULATOR-2*. Universitas Islam Negeri Syarif Hidayatullah Jakarta.
- Sidik, B. (2014). *Pemrograman Web dengan PHP* (2nd ed.). Bandung: Informatika Bandung.
- Simarangkir, M. (2017). STUDI PERBANDINGAN ALGORITMA - ALGORITMA STEMMING UNTUK DOKUMEN TEKS BAHASA INDONESIA, 1. Retrieved from www.politeknikmeta.ac.id/meta/ojs/index.php/inkofar/article/download/2/2
- Sjukani, M. (2014). *ALGORITMA : [Algoritma dan Struktur Data 1] dengan C,*

C++, dan Java-- Teknik-Teknik Dasar Pemrograman Komputer. Jakarta: Mitra Wacana Media.

Subari, & Ferdinandus. (2015). SISTEM INFORMATION RETRIEVAL LAYANAN KESEHATAN UNTUK BEROBAT DENGAN METODE VECTOR SPACE MODEL (VSM) BERBASIS WEBGIS. Retrieved from https://www.researchgate.net/publication/286334182_SISTEM_INFORMATION_RETRIEVAL_LAYANAN_KESEHATAN_UNTUK_BEROBAT_DENGAN_METODE_VECTOR_SPACE_MODEL_VSM_BERBASIS_WEBGIS

Wahyudi, D., Susyanto, T., & Nugroho, D. (2017). IMPLEMENTASI DAN ANALISIS ALGORITMA STEMMING NAZIEF & ADRIANI DAN PORTER PADA DOKUMEN BERBAHASA INDONESIA. Retrieved from p3m.sinus.ac.id/jurnal/index.php/e-jurnal_SINUS/article/download/305/pdf

Widodo, & Wahyuni, D. (2017). IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK MENGETAHUI BIDANG SKRIPSI MAHASISWA MULTIMEDIA PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER UNIVERSITAS NEGERI JAKARTA, *1 No. 2*. Retrieved from https://www.researchgate.net/publication/327466558_MPLEMENTASI_ALGORITMA_K-MEANS_CLUSTERING_UNTUK_MENGETAHUI_BIDANG_SKRIPSI_MAHASISWA_MULTIMEDIA_PENDIDIKAN_TEKNIK_INFORMATIKA_DAN_KOMPUTER_UNIVERSITAS_NEGERI_JAKARTA

LAMPIRAN

Lampiran I. Wawancara

Pewawancara : Muhammad Haris Al Farisi

Hari, tanggal : Rabu, 24 Oktober 2018

Waktu : 13.30 s/d selesai

Narasumber : Any Sayekti dan Sopha Julia

Tempat : Kemendikbud RI

1. Pertanyaan : Kapan sistem zonasi sekolah berjalan dan apakah sudah dilaksanakan di seluruh Indonesia?

Jawaban:

Sistem Zonasi Sekolah ditetapkan tanggal 2 Mei 2018 dan diundangkan pada 8 Mei 2018.

2. Pertanyaan : Apa tujuan dilaksanakannya sistem zonasi?

Jawaban:

Menurut Kemendikbud sistem zonasi untuk mendekatkan jarak tempuh peserta didik ke sekolah namun menurut PDSPK (Pusat Data Statistik Pendidikan dan Kebudayaan) sebaliknya misalnya ada satu pusat sekolah yang menjadi zonasi.

3. Pertanyaan : Apa saja masalah yang terjadi selama pelaksanaan sistem zonasi?

Jawaban:

Sistem zonasi sekolah memiliki beberapa masalah diantaranya yaitu waktu sosialisasi yang mepet dengan pelaksanaan PPDB tahun 2018. Dalam sosialisasi ke kabupaten kota provinsi yang hadir tidak semuanya menyampaikan ke masyarakat ataupun *stakeholder* yang ada dibawahnya. Kekurangpahaman pemda dalam memahami dan menerjemahkan dalam bentuk jenis itu di daerahnya. Masih banyak daerah-daerah yang tidak

sesuai dengan peraturan nomor 14 tahun 2018 seperti penerimaan PPDB yang tidak semua menerapkan penerimaan minimal 90% siswa terdekat dari rumah ke sekolah.

4. Pertanyaan : Apakah sistem zonasi ini memiliki masa berlaku?

Jawaban:

Tidak memiliki masa berlaku namun kebijakan yang sedang berjalan ini masih menerima evaluasi dan menentukan konsep kedepannya agar sistem dapat terus berjalan berdasarkan evaluasi PPDB tiap tahun. Dalam meningkatkan mutu pendidikan dapat dilakukan melalui intervensi kebijakan.

