

TUGAS AKHIR

FUZZY CLUSTERING UNTUK TEXT DIMENSIONALITY REDUCTION



Jeremy Christian Budiawan

NPM: 6181901017

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2024

FINAL PROJECT

**FUZZY CLUSTERING FOR TEXT DIMENSIONALITY
REDUCTION**



Jeremy Christian Budiawan

NPM: 6181901017

**DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2024**

LEMBAR PENGESAHAN

**FUZZY CLUSTERING UNTUK TEXT DIMENSIONALITY
REDUCTION**

Jeremy Christian Budiawan

NPM: 6181901017

Bandung, 25 Juni 2024

Menyetujui,

Pembimbing

**Digitally signed
by Husnul
Hakim**

Husnul Hakim, M.T.

Ketua Tim Penguji

**Digitally signed
by Luciana
Abednego**

Luciana Abednego, M.T.

Anggota Tim Penguji

**Digitally signed
by Raymond
Chandra Putra**

Raymond Chandra Putra, M.T.

Mengetahui,

Ketua Program Studi

**Digitally signed
by Lionov**

Lionov, Ph.D.

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa tugas akhir dengan judul:

FUZZY CLUSTERING UNTUK TEXT DIMENSIONALITY REDUCTION

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 25 Juni 2024



Jeremy Christian Budiawan
NPM: 6181901017

ABSTRAK

Digitalisasi membawa pengaruh besar dalam keberlangsungan hidup manusia. Cara manusia mendapatkan informasi berubah, dari yang awalnya mengandalkan media cetak seperti koran dan majalah, kini beralih ke media digital seperti situs web dan aplikasi. Pemrosesan teks atau analisis data teks menjadi semakin penting, memungkinkan kita untuk mengekstrak informasi berharga dari berbagai sumber digital. Namun, data teks perlu dimodelkan sedemikian rupa sehingga dapat diproses oleh komputer. Pemodelan data teks menjadi vektor dalam ruang berdimensi, memiliki permasalahan yaitu *high dimensional* dan *sparsity*.

Beberapa teknik reduksi dimensi yang cukup umum antara lain, *Principal Component Analysis* (PCA), *Latent Semantic Analysis* (LSA) dan *t-Distributed Stochastic Neighbor Embedding* (t-SNE). Namun pada penelitian ini dilakukan eksperimen implementasi reduksi dimensi dengan teknik klusterisasi, pendekatan *fuzzy clustering* (*soft clustering*). *Fuzzy clustering* memungkinkan suatu objek menjadi anggota pada lebih dari satu kelompok (klaster) dengan persentase atau derajat keanggotaan yang berbeda pada setiap kelompoknya. Pendekatan ini dapat membantu representasi yang lebih baik dari data teks, karena kata-kata yang terkandung dalam data bisa serupa untuk berbagai topik. Interpretasi dimensi yang awalnya adalah kata unik dalam kumpulan dokumen diubah menjadi derajat keanggotaan pada klaster. Sehingga membuat *dataset* tidak bersifat *sparse* dan dimensinya lebih kecil.

Data eksperimen yang digunakan merupakan data ulasan film, program televisi, dan video lainnya. Data ulasan film dari platform online IMDb didapatkan dari Kaggle. *Dataset* ini merupakan data yang digunakan oleh peneliti-peneliti di program studi dan penelitian kecerdasan buatan (*artificial intelligence*) di Universitas Stanford, Andrew L. Maas dan kawan-kawan.

Berdasarkan eksperimen yang telah dilakukan, diperoleh bahwa reduksi dimensi dengan pendekatan *fuzzy clustering* berhasil mengurangi jumlah dimensi pada data dan mengatasi permasalahan *sparsity*. Selain itu dilakukan perbandingan performa model klasifikasi yang dilatih menggunakan data yang direduksi dan tidak direduksi. Hasil eksperimen menunjukkan bahwa performa kedua model klasifikasi tidak memiliki perbedaan yang signifikan. Namun, perlu dilakukan penelitian lebih lanjut untuk memastikan hasil yang lebih optimal.

Kata-kata kunci: Data Teks, Reduksi Dimensi, *Fuzzy Clustering*

ABSTRACT

Digitalization has significantly impacted human life. The way people obtain information has changed, from relying on printed media such as newspapers and magazines to digital media like websites and applications. Text processing or text data analysis has become increasingly important, allowing us to extract valuable information from various digital sources. Text data needs to be modeled in such a way that it can be processed by computers. Modeling text data into vectors in dimensional space has problems, namely high dimensionality and sparsity.

Some common dimensionality reduction techniques include Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and t-Distributed Stochastic Neighbor Embedding (t-SNE). However, in this research, an experiments were conducted to implement dimensionality reduction using clustering techniques, specifically the fuzzy clustering (soft clustering) approach. Fuzzy clustering allows an object to belong to more than one group (cluster) with different percentages or degrees of membership in each group. This approach can help provide a better representation of text data, as the words contained in the data can be similar for various topics. The interpretation of dimensions, which initially are unique words in the document collection, is transformed into membership degrees in clusters. Thus, it makes the dataset not sparse and has smaller dimensions.

The experimental data used includes film reviews, TV programs, and other video reviews. Film review data from the online platform IMDb was obtained from Kaggle. This dataset is used by researchers in the artificial intelligence program and research at Stanford University, Andrew L. Maas, and colleagues.

Based on the experiments, it was found that dimension reduction using the fuzzy clustering approach successfully reduced the number of dimensions in the data and addressed the sparsity problem. Additionally, a performance comparison was made between classification models trained using reduced and non-reduced data. The experimental results showed that the performance of both classification models did not have significant differences. However, further research is needed to ensure more optimal results.

Keywords: Text Data, Dimension Reduction, Fuzzy Clustering

Dipersembahkan untuk kedua orang tua...

KATA PENGANTAR

Puji dan syukur Penulis panjatkan kepada Tuhan yang Maha Esa, atas berkat dan rahmat-Nya penulis dapat menyelesaikan Tugas Akhir yang berjudul “Fuzzy Clustering untuk Text Dimensionality Reduction”. Tugas Akhir ini dibuat sebagai salah satu syarat untuk menyelesaikan pendidikan di Program Studi Informatika, Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan, Bandung.

Dalam penyusunan dokumen Tugas Akhir ini tak lupa peneliti memberikan ucapan terima kasih kepada pihak-pihak yang terlibat baik dalam memberi bimbingan, bantuan dan dukungan secara langsung atau tidak ke dalam proses penyusunan penelitian, terutama untuk:

1. Kepada kedua orang tua, adik dan kakak Penulis yang selalu mendukung, menyemangati dan memberikan doa.
2. Kepada Ibu Maria Veronica Claudia, M.T. selaku dosen pembimbing, yang telah banyak memberikan ilmu, saran, dan bantuan selama proses pengerjaan tugas akhir ini.
3. Kepada Ibu Luciana Abednego, M.T. dan Bapak Raymond Chandra Putra, M.T. selaku dosen penguji yang telah memberikan masukan dan saran untuk Tugas Akhir ini.
4. Kepada teman-teman yang menemani Penulis selama masa kuliah, Axel, Daryl, Mike, Mone, Ferell, dan Indra.
5. Kepada seluruh pihak yang mendukung Penulis yang tidak dapat disebutkan satu-satu

Penulis juga ingin berterima kasih kepada pembaca yang telah membaca Tugas Akhir ini, Penulis berharap Tugas Akhir ini dapat bermanfaat dan membantu untuk kemajuan ilmu pengetahuan, dan dapat dijadikan panduan untuk penelitian selanjutnya

Bandung, Juni 2024

Penulis

DAFTAR ISI

KATA PENGANTAR	xv
DAFTAR ISI	xvii
DAFTAR GAMBAR	xxi
DAFTAR TABEL	xxv
DAFTAR KODE PROGRAM	xxix
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	5
1.3 Tujuan	6
1.4 Batasan Masalah	6
1.5 Metodologi	6
1.6 Sistematika Pembahasan	6
2 DASAR TEORI	9
2.1 <i>Natural Language Processing</i>	9
2.1.1 <i>Vector Space Model</i>	10
2.1.2 <i>Document Term Matrix</i>	10
2.1.3 TF-IDF	11
2.1.4 <i>Regular Expression (Regex)</i>	12
2.1.5 <i>Part of Speech Tagging (PoS) Tagging</i>	13
2.1.6 <i>Text Preprocessing</i>	13
2.2 <i>Machine Learning</i>	15
2.3 <i>Supervised Learning</i>	16
2.3.1 Klasifikasi	16
2.3.2 <i>Naïve Bayes Classifier</i>	16
2.3.3 <i>Decision Tree Classifier</i>	18
2.3.4 <i>Random Forest Classifier</i>	20
2.3.5 <i>Logistic Regression</i>	21
2.3.6 <i>K Nearest Neighbors (K-NN)</i>	23
2.4 <i>Unsupervised Learning</i>	24
2.4.1 Klasterisasi	24
2.4.2 <i>Hard Clustering</i>	25
2.4.3 <i>Fuzzy Clustering</i>	26
2.5 <i>Distance Metrics</i>	28
2.6 Indeks validasi hasil klasterisasi	28
2.6.1 <i>Davies Bouldin Index</i>	29
2.6.2 <i>Partition Coefficient</i>	29
2.7 Indeks validasi hasil klasifikasi	30

2.7.1	<i>Confusion Matrix</i>	30
2.7.2	<i>Accuracy</i>	31
2.7.3	<i>Precision</i>	31
2.7.4	<i>Recall</i>	31
2.7.5	<i>F1-score</i>	31
2.8	Reduksi Dimensi	31
2.9	Transformasi Data melalui Normalisasi	32
2.10	Visualisasi Data	33
2.11	<i>Library Python</i>	35
3	ANALISIS PENYELESAIAN MASALAH	37
3.1	Deskripsi Masalah	37
3.2	Tahap Penyelesaian Masalah	38
3.3	Eksperimen Vektorisasi TF-IDF	39
3.3.1	Library TfidfVectorizer	39
3.3.2	Komputasi Manual TF-IDF	41
3.4	Eksperimen Klasterisasi dengan Fuzzy C-Means	43
3.4.1	Library cmeans	43
3.4.2	Komputasi Manual fuzzy-cmeans	48
4	PENAMBANGAN DATA	57
4.1	Pengumpulan Data	57
4.2	Penyiapan Data	62
4.3	Eksplorasi Data	65
4.4	Analisis Data	72
4.4.1	Eksperimen Pertama Pembuatan Model Analisis Sentimen	73
4.4.2	Eksperimen Kedua Pembangunan Model Machine Learning dan Evaluasi	75
4.4.3	Eksperimen Ketiga Pembangunan Model Machine Learning dan Evaluasi	79
4.4.4	Eksperimen Keempat Pembangunan Model Machine Learning dan Evaluasi	84
4.4.5	Eksperimen Kelima Pembangunan Model Machine Learning dan Evaluasi	93
4.5	Kesimpulan	95
5	PELUNCURAN MODEL DAN PENGUJIAN	97
5.1	Pembangunan Perangkat Lunak	97
5.1.1	Fitur Perangkat Lunak	97
5.1.2	Diagram <i>Use Case</i>	97
5.2	Perancangan Tampilan Antarmuka	99
5.2.1	Halaman <i>home</i>	99
5.2.2	Halaman prediksi	100
5.3	Implementasi Perangkat Lunak	100
5.3.1	Halaman Utama	101
5.3.2	Halaman Prediksi	101
5.3.3	Halaman Hasil Prediksi	102
5.4	Pengujian Fungsionalitas Perangkat Lunak	103
6	KESIMPULAN DAN SARAN	105
6.1	Kesimpulan	105
6.2	Saran	105
	DAFTAR REFERENSI	107
	A KODE PROGRAM	109

B	HASIL EKSPERIMEN	119
B.1	Iterasi Pertama Pembuatan Model Analisis Sentimen	119
B.2	Iterasi Kedua Pembuatan Model Analisis Senitmen	120
B.2.1	Soft Clustering	120
B.3	Iterasi Ketiga Pembuatan Model Analisis Senitmen	168
B.3.1	Soft Clustering	168
B.4	Iterasi Keempat Pembuatan Model Analisis Sentimen	169
B.4.1	Soft Clustering	169
B.5	Iterasi Kelima Pembuatan Model Analisis Sentimen	171

DAFTAR GAMBAR

1.1	Jumlah pengguna internet di Indonesia	1
1.2	Pemodelan data teks dalam ruang tiga dimensi	2
1.3	<i>Hard clustering</i> dan <i>fuzzy clustering</i>	3
1.4	Titik pusat <i>cluster</i>	4
1.5	Reduksi dimensi dengan <i>fuzzy clustering</i>	4
1.6	Reduksi dimensi dengan <i>hard clustering</i>	5
1.7	Alur proses pengerjaan TA	5
2.1	<i>Vector Space Model</i>	10
2.2	<i>Document Term Matrix</i>	11
2.3	<i>Document Term Matrix</i>	11
2.4	<i>Part of Speech (POS) Tag</i>	13
2.5	<i>Part of Speech (POS) Tag</i>	13
2.6	<i>Machine learning</i>	15
2.7	Diagram <i>Decision Tree</i>	19
2.8	<i>Random Forest</i>	21
2.9	<i>K-NN</i>	23
2.10	Klasterisasi	24
2.11	<i>Hard clustering</i>	25
2.12	Alur <i>K-Means</i>	25
2.13	<i>fuzzy clustering</i>	26
2.14	Alur <i>FCM</i>	28
2.15	<i>Confusion Matrix</i>	30
2.16	Contoh <i>feature extraction</i>	32
2.17	Contoh <i>feature selection</i>	32
2.18	Visualisasi <i>bar chart</i> kategori dengan jumlah	33
2.19	Visualisasi <i>line chart</i> perkembangan harga BTC	34
2.20	Visualisasi <i>word cloud</i> dari komentar-komentar berbagai film	34
3.1	Reduksi dimensi dengan <i>fuzzy clustering</i> [24]	38
3.2	Reduksi dimensi dengan <i>hard clustering</i>	39
3.3	Data review	40
3.4	Hasil Matriks TF-IDF	41
3.5	Hasil Kosa Kata	41
4.1	File yang diunduh	57
4.2	<i>File</i> yang telah diekstrak	57
4.3	Isi <i>folder</i> train	58
4.4	Isi <i>folder</i> neg	58
4.5	<i>File</i> .txt yang bersentimen negatif	59
4.6	Gabungan <i>file</i> ulasan bersentimen negatif	59
4.7	Potongan <i>file</i> ulasan yang telah digabungkan	60
4.8	Data duplikat	62

4.9	Distribusi jumlah kata pada <i>dataset</i>	65
4.10	<i>Word cloud dataset</i>	67
4.11	<i>Word cloud dataset</i> bersentimen positif	68
4.12	<i>Word cloud dataset</i> bersentimen negatif	68
4.13	<i>Bar chart</i> jumlah kata sifat terbanyak	70
4.14	<i>Bar chart</i> kata sifat terbanyak pada dataframe bersentimen positif	71
4.15	<i>Bar chart</i> kata sifat terbanyak pada dataframe bersentimen negatif	71
4.16	Potongan dokumen dalam bentuk <i>dataframe</i>	73
4.17	Potongan hasil vektorisasi TF-IDF dokumen	73
4.18	<i>Line chart</i> nilai akurasi model ML	74
4.19	Potongan dokumen dalam bentuk <i>dataframe</i>	75
4.20	Potongan hasil vektorisasi TF-IDF dokumen	76
4.21	Bentuk data hasil reduksi dimensi 2 klaster - FCM	76
4.22	<i>Line chart</i> nilai akurasi model ML	77
4.23	Bentuk data hasil reduksi dimensi 1216 klaster - K-Means	78
4.24	<i>Line chart</i> nilai DB	81
4.25	Potongan bentuk data hasil reduksi dimensi 1992 klaster - FCM	81
4.26	Perbandingan <i>accuracy</i> dengan jumlah dimensi 1992	82
4.27	Potongan bentuk data hasil reduksi dimensi 1992 klaster - K-Means	83
4.28	Potongan bentuk data hasil reduksi dimensi 614 klaster - FCM	85
4.29	Perbandingan <i>accuracy</i> dengan jumlah dimensi 614	86
4.30	Potongan bentuk data hasil reduksi dimensi 687 klaster - FCM	86
4.31	Perbandingan <i>accuracy</i> dengan jumlah dimensi 687	87
4.32	Potongan bentuk data hasil reduksi dimensi 979 klaster - FCM	88
4.33	Perbandingan <i>accuracy</i> dengan jumlah dimensi 979	89
4.34	Potongan bentuk data hasil reduksi dimensi 1366 klaster - FCM	89
4.35	Perbandingan <i>accuracy</i> dengan jumlah dimensi 1366	90
4.36	Potongan bentuk data hasil reduksi dimensi 1366 klaster - K-Means	91
4.37	Potongan bentuk data hasil reduksi dimensi 1216 klaster - FCM	93
4.38	Potongan bentuk data hasil <i>min-max scaling</i>	93
5.1	Diagram <i>use case</i>	99
5.2	Rancangan tampilan halaman <i>home</i>	99
5.3	Rancangan tampilan halaman prediksi data	100
5.4	Tampilan halaman utama pada <i>website</i>	101
5.5	Tampilan halaman untuk prediksi	101
5.6	Tampilan halaman hasil prediksi	102
5.7	Tampilan halaman hasil prediksi, <i>show chart</i>	102
5.8	Tampilan halaman hasil prediksi, <i>pagination</i>	103
B.1	Perbandingan <i>accuracy</i> dengan jumlah dimensi 2	120
B.2	Perbandingan <i>accuracy</i> dengan jumlah dimensi 2	121
B.3	Perbandingan <i>accuracy</i> dengan jumlah dimensi 2	122
B.4	Perbandingan <i>accuracy</i> dengan jumlah dimensi 2	123
B.5	Perbandingan <i>accuracy</i> dengan jumlah dimensi 20	124
B.6	Perbandingan <i>accuracy</i> dengan jumlah dimensi 20	125
B.7	Perbandingan <i>accuracy</i> dengan jumlah dimensi 20	126
B.8	Perbandingan <i>accuracy</i> dengan jumlah dimensi 20	127
B.9	Perbandingan <i>accuracy</i> dengan jumlah dimensi 101	128
B.10	Perbandingan <i>accuracy</i> dengan jumlah dimensi 101	129
B.11	Perbandingan <i>accuracy</i> dengan jumlah dimensi 101	130
B.12	Perbandingan <i>accuracy</i> dengan jumlah dimensi 202	131

B.13 Perbandingan <i>accuracy</i> dengan jumlah dimensi 202	132
B.14 Perbandingan <i>accuracy</i> dengan jumlah dimensi 202	133
B.15 Perbandingan <i>accuracy</i> dengan jumlah dimensi 202	134
B.16 Perbandingan <i>accuracy</i> dengan jumlah dimensi 202	135
B.17 Perbandingan <i>accuracy</i> dengan jumlah dimensi 405	136
B.18 Perbandingan <i>accuracy</i> dengan jumlah dimensi 405	137
B.19 Perbandingan <i>accuracy</i> dengan jumlah dimensi 405	138
B.20 Perbandingan <i>accuracy</i> dengan jumlah dimensi 405	139
B.21 Perbandingan <i>accuracy</i> dengan jumlah dimensi 608	140
B.22 Perbandingan <i>accuracy</i> dengan jumlah dimensi 608	141
B.23 Perbandingan <i>accuracy</i> dengan jumlah dimensi 608	142
B.24 Perbandingan <i>accuracy</i> dengan jumlah dimensi 608	143
B.25 Perbandingan <i>accuracy</i> dengan jumlah dimensi 810	144
B.26 Perbandingan <i>accuracy</i> dengan jumlah dimensi 810	145
B.27 Perbandingan <i>accuracy</i> dengan jumlah dimensi 810	146
B.28 Perbandingan <i>accuracy</i> dengan jumlah dimensi 810	147
B.29 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1013	148
B.30 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1013	149
B.31 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1013	150
B.32 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1013	151
B.33 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1216	152
B.34 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1216	153
B.35 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1216	154
B.36 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1216	155
B.37 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1419	156
B.38 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1419	157
B.39 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1419	158
B.40 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1419	159
B.41 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1621	160
B.42 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1621	161
B.43 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1621	162
B.44 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1621	163
B.45 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1824	164
B.46 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1824	165
B.47 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1824	166
B.48 Perbandingan <i>accuracy</i> dengan jumlah dimensi 1824	167

DAFTAR TABEL

1.1	Contoh <i>Document Term Matrix</i> (DTM)	2
2.1	DTM TF-IDF	12
2.2	Contoh <i>meta character</i> dengan kegunaannya	12
2.3	Contoh dari 3 <i>regex function</i>	12
2.4	Teks sebelum dan setelah dilakukan <i>cleaning</i>	14
2.5	Teks sebelum dan setelah dilakukan <i>case folding</i>	14
2.6	Teks sebelum dan setelah dilakukan penghapusan <i>stopwords</i>	14
2.7	Teks sebelum dan setelah dilakukan <i>word tokenization</i>	14
2.8	Teks sebelum dan setelah dilakukan <i>lemmatization</i>	15
2.9	<i>Dataset supervised learning</i>	16
2.10	Jumlah Guru dan Murid	17
2.11	<i>Dataset unsupervised learning</i>	24
2.12	Tabel perbandingan nilai DB	29
3.1	Nilai TF Dokumen 1	42
3.2	Nilai TF Dokumen 2	42
3.3	Nilai TF Dokumen 3	42
3.4	Nilai TF Dokumen 4	42
3.5	Nilai TF Dokumen 1	42
3.6	TF-IDF Dokumen 1	42
3.7	TF-IDF Dokumen 2	42
3.8	TF-IDF Dokumen 3	42
3.9	TF-IDF Dokumen 4	42
3.10	DTM eksperimen perhitungan TF-IDF	43
3.11	Tabel parameter algoritma FCM	43
3.12	Tabel hasil algoritma FCM	44
3.13	Rincian data	44
3.14	<i>Dataset</i>	44
3.15	<i>Dataset</i> TF-IDF	45
3.16	<i>Dataset</i> hasil reduksi dimensi	46
3.17	Titik pusat kluster	46
3.18	Derajat keanggotan antar kluster	46
3.19	Derajat keanggotan antar kluster awal	47
3.20	Jarak antara data dengan titik pusat kluster	47
3.21	Riwayat nilai FO	47
3.22	Nilai p eksperimen 1	47
3.23	Nilai fpc eksperimen ke-1	47
3.24	Normalisasi fitur	48
3.25	<i>Dataset</i> eksperimen	48
3.26	<i>Dataset</i> vektorisasi TF-IDF	48
3.27	Nilai μ_{ik}^m	49
3.28	Nilai $(\mu_{ik}^m)^m \times$ data sampel ke- i variabel ke- j kluster 1	50

3.29	Nilai $(\mu_{ik})^m \times$ data sampel ke- i variabel ke- j klaster 2	50
3.30	Pusat klaster iterasi ke-1	50
3.31	Total $(x_{ij} - v_{kj})^2$ klaster 1	50
3.32	Perhitungan fungsi objektif klaster 1 iterasi 1	51
3.33	Total $(x_{ij} - v_{kj})^2$ klaster 2	51
3.34	Perhitungan fungsi objektif klaster 2 iterasi 1	51
3.35	Hasil akhir perhitungan fungsi objektif iterasi ke-1	51
3.36	Nilai partisi U yang baru	52
3.37	Iterasi pertama fungsi objektif	52
3.38	Nilai μ_{ik}^m	52
3.39	Nilai $(\mu_{ik})^m \times$ data sampel ke- i variabel ke- j klaster 1	53
3.40	Nilai $(\mu_{ik})^m \times$ data sampel ke- i variabel ke- j klaster 2	53
3.41	Pusat klaster iterasi ke-2	53
3.42	Total $(x_{ij} - v_{kj})^2$ klaster 1	53
3.43	Perhitungan fungsi objektif klaster 1 iterasi 2	54
3.44	Total $(x_{ij} - v_{kj})^2$ klaster 2	54
3.45	Perhitungan fungsi objektif klaster 2 iterasi 2	54
3.46	Hasil akhir perhitungan fungsi objektif iterasi ke-2	54
3.47	Nilai partisi U yang baru	55
3.48	Iterasi kedua fungsi objektif	55
3.49	Derajat keanggotaan data untuk setiap klaster	55
4.1	<i>Dataset</i>	61
4.2	Data sebelum dan sesudah melalui penyiapan data	64
4.3	10 kata umum yang terdapat pada <i>dataset</i> bersentimen negatif	69
4.4	10 kata umum yang terdapat pada <i>dataset</i> bersentimen positif	69
4.5	Confusion Matrix	74
4.6	Hasil evaluasi model Multinomial Naive Bayes	74
4.7	Eksperimen penurunan nilai dimensi	75
4.8	<i>Confusion matrix</i> model K-NN	77
4.9	Hasil evaluasi model K-NN	77
4.10	<i>Confusion matrix</i> model K-NN	78
4.11	Hasil evaluasi model K-NN	78
4.12	Perbandingan hasil evaluasi ML	79
4.13	Nilai DB dari jumlah klaster sebanyak 2–1999	80
4.14	<i>Confusion matrix</i> model MLR	82
4.15	Hasil evaluasi model K-NN	82
4.16	Confusion Matrix	83
4.17	Hasil evaluasi model KNN	83
4.18	Perbandingan hasil evaluasi ML	84
4.19	Hasil evaluasi model	84
4.20	<i>Confusion matrix</i> model MLR	85
4.21	Hasil evaluasi model K-NN	85
4.22	<i>Confusion matrix</i> model K-NN	87
4.23	Hasil evaluasi model K-NN	87
4.24	<i>Confusion matrix</i> model K-NN	88
4.25	Hasil evaluasi model K-NN	88
4.26	<i>Confusion matrix</i> model K-NN	90
4.27	Hasil evaluasi model K-NN	90
4.28	Hasil evaluasi model K-NN	91
4.29	Confusion Matrix	92
4.30	Hasil evaluasi model KNN	92

4.31 Hasil evaluasi model K-NN	92
4.32 <i>Confusion matrix</i> model LR	94
4.33 Hasil evaluasi model LR dengan Train/Test	94
4.34 Hasil evaluasi model dengan Train/Test	94
4.35 Perbandingan model ML tanpa reduksi dimensi dan dengan reduksi dimensi	96
5.1 Skenario pengguna membuka <i>website</i>	97
5.2 Skenario pengguna klik tombol “click to start”	98
5.3 Skenario pengguna melakukan prediksi	98
5.4 Skenario pengguna mengunduh hasil prediksi	98
5.5 Pengujian dan hasilnya	103
5.6 Potongan data uji	104
B.1 <i>Classification report</i> iterasi pertama	119
B.2 <i>Classification report</i> iterasi kedua, parameter pertama	120
B.3 <i>Classification report</i> iterasi kedua, parameter kedua	121
B.4 <i>Classification report</i> iterasi kedua, parameter ketiga	122
B.5 <i>Classification report</i> iterasi kedua, parameter keempat	123
B.6 <i>Classification report</i> iterasi kedua, parameter kelima	124
B.7 <i>Classification report</i> iterasi kedua, parameter keenam	125
B.8 <i>Classification report</i> iterasi kedua, parameter ketujuh	126
B.9 <i>Classification report</i> iterasi kedua, parameter kedelapan	127
B.10 <i>Classification report</i> iterasi kedua, parameter kesembilan	128
B.11 <i>Classification report</i> iterasi kedua, parameter kesepuluh	129
B.12 <i>Classification report</i> iterasi kedua, parameter kesebelas	130
B.13 <i>Classification report</i> iterasi kedua, parameter keduabelas	131
B.14 <i>Classification report</i> iterasi kedua, parameter ketigabelas	132
B.15 <i>Classification report</i> iterasi kedua, parameter keempatbelas	133
B.16 <i>Classification report</i> iterasi kedua, parameter kelimabelas	134
B.17 <i>Classification report</i> iterasi kedua, parameter keenambelas	135
B.18 <i>Classification report</i> iterasi kedua, parameter ketujuhbelas	136
B.19 <i>Classification report</i> iterasi kedua, parameter kedelapanbelas	137
B.20 <i>Classification report</i> iterasi kedua, parameter kesembilabelas	138
B.21 <i>Classification report</i> iterasi kedua, parameter keduapuluh	139
B.22 <i>Classification report</i> iterasi kedua, parameter ke-21	140
B.23 <i>Classification report</i> iterasi kedua, parameter ke-22	141
B.24 <i>Classification report</i> iterasi kedua, parameter ke-23	142
B.25 <i>Classification report</i> iterasi kedua, parameter ke-24	143
B.26 <i>Classification report</i> iterasi kedua, parameter ke-25	144
B.27 <i>Classification report</i> iterasi kedua, parameter ke-26	145
B.28 <i>Classification report</i> iterasi kedua, parameter ke-27	146
B.29 <i>Classification report</i> iterasi kedua, parameter ke-28	147
B.30 <i>Classification report</i> iterasi kedua, parameter ke-29	148
B.31 <i>Classification report</i> iterasi kedua, parameter ke-30	149
B.32 <i>Classification report</i> iterasi kedua, parameter ke-31	150
B.33 <i>Classification report</i> iterasi kedua, parameter ke-32	151
B.34 <i>Classification report</i> iterasi kedua, parameter ke-33	152
B.35 <i>Classification report</i> iterasi kedua, parameter ke-34	153
B.36 <i>Classification report</i> iterasi kedua, parameter ke-35	154
B.37 <i>Classification report</i> iterasi kedua, parameter ke-36	155
B.38 <i>Classification report</i> iterasi kedua, parameter ke-37	156
B.39 <i>Classification report</i> iterasi kedua, parameter ke-38	157

B.40	<i>Classification report</i> iterasi kedua, parameter ke-39	158
B.41	<i>Classification report</i> iterasi kedua, parameter ke-40	159
B.42	<i>Classification report</i> iterasi kedua, parameter ke-41	160
B.43	<i>Classification report</i> iterasi kedua, parameter ke-42	161
B.44	<i>Classification report</i> iterasi kedua, parameter ke-43	162
B.45	<i>Classification report</i> iterasi kedua, parameter ke-44	163
B.46	<i>Classification report</i> iterasi kedua, parameter ke-45	164
B.47	<i>Classification report</i> iterasi kedua, parameter ke-46	165
B.48	<i>Classification report</i> iterasi kedua, parameter ke-47	166
B.49	<i>Classification report</i> iterasi kedua, parameter ke-48	167
B.50	<i>Classification report</i> iterasi ketiga	168
B.51	<i>Classification report</i> iterasi keempat, parameter ke-1	169
B.52	<i>Classification report</i> iterasi keempat, parameter ke-2	169
B.53	<i>Classification report</i> iterasi keempat, parameter ke-3	170
B.54	<i>Classification report</i> iterasi keempat, parameter ke-4	170
B.55	<i>Classification report</i> iterasi kelima	171

DAFTAR KODE PROGRAM

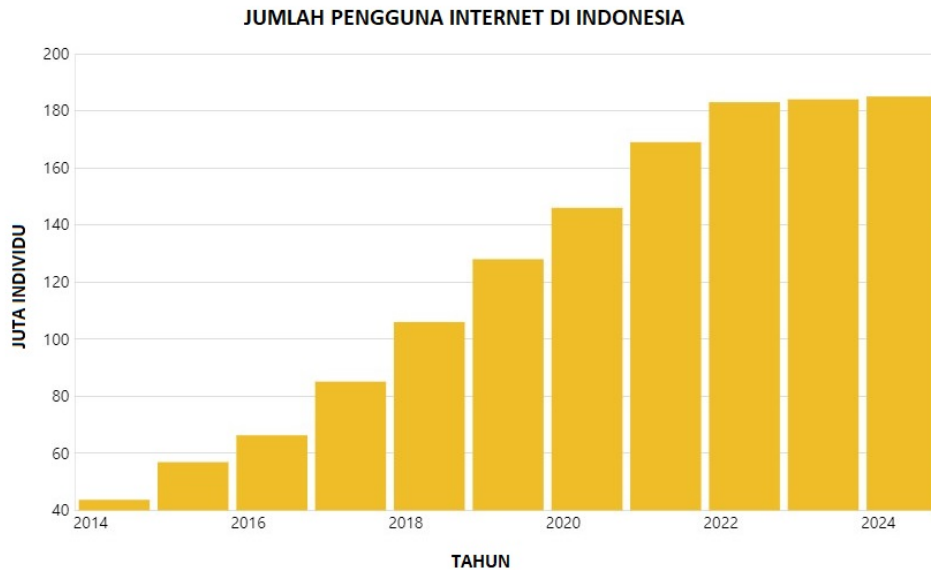
3.1	Kode untuk membuat data review	40
3.2	Kode untuk merubah teks menjadi vektor TF-IDF	40
3.3	Kode untuk menampilkan kata	41
3.4	Kode untuk implementasi FCM	45
3.5	Kode untuk melakukan <i>fuzzy clustering</i>	45
3.6	Kode untuk menyimpan hasil pengelompokkan	45
4.1	Kode untuk menggabungkan file bersentimen negatif dan positif	59
4.2	Kode untuk menggabungkan file bersentimen negatif dengan positif	60
4.3	Kode untuk memotong dataset	62
4.4	Kode untuk mengecek <i>null values</i>	62
4.5	Kode untuk menghapus data duplikat	62
4.6	Kode untuk membersihkan teks dari tanda baca	63
4.7	Kode untuk membersihkan teks dari elemen HTML	63
4.8	Kode untuk menghapus spasi yang berlebihan	63
4.9	Kode untuk melakukan tokenisasi kata	63
4.10	Kode untuk merubah seluruh huruf menjadi huruf kecil	63
4.11	Kode untuk menghapus kata yang termasuk ke dalam stop word	63
4.12	Kode untuk merubah kata ke bentuk kata dasarnya	63
4.13	Kode untuk visualisasi histogram distribusi persebaran jumlah kata	66
4.14	Kode untuk mencari nilai rata-rata kemunculan kata	66
4.15	Kode untuk mencari nilai median kemunculan kata	66
4.16	Kode untuk mencari jumlah kata unik	66
4.17	Kode untuk menampilkan <i>word cloud</i>	66
4.18	Kode untuk menampilkan <i>word cloud dataset</i> bersentimen positif dan negatif	67
4.19	Kode untuk mencari jumlah kata sifat terbanyak dan visualisasinya	69
4.20	Kode untuk menampilkan kata sifat pada data bersentimen positif	70
A.1	Kode penyiapan data	109
A.2	Kode untuk mengevaluasi hasil pengelompokkan berdasarkan nilai DB	110
A.3	Kode untuk menampilkan line chart nilai DB	111
A.4	Kode untuk melakukan klasterisasi dengan algoritma FCM	111
A.5	Kode untuk melakukan klasterisasi dengan algoritma K-means	111
A.6	Kode untuk melakukan transformasi min-max	112
A.7	Kode untuk melakukan klasifikasi dataset reduksi dimensi FCM	112
A.8	Kode untuk melakukan klasifikasi dataset reduksi dimensi K-means	114
A.9	Kode untuk melakukan klasifikasi dataset reduksi tambah normalisasi	115
A.10	Kode untuk melakukan train model dan save model	117

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Teknologi komputer dan internet telah merevolusi cara manusia menjalani kehidupan¹. Komputer dengan kemampuannya yang luar biasa, menjadi elemen penting dalam berbagai bidang, mulai dari bisnis hingga hiburan. Internet sebagai jaringan global, menghubungkan jutaan orang di seluruh dunia dan mengubah cara manusia untuk berkomunikasi, mencari informasi serta berkolaborasi. Sampai bulan Januari tahun 2024 terdapat 180 juta pengguna internet. Penambahan pengguna internet di Indonesia pada 10 tahun terakhir dapat dilihat pada Gambar 1.1². Dari tahun 2014 sampai tahun 2024 telah terjadi penambahan pengguna internet yang sangat signifikan.



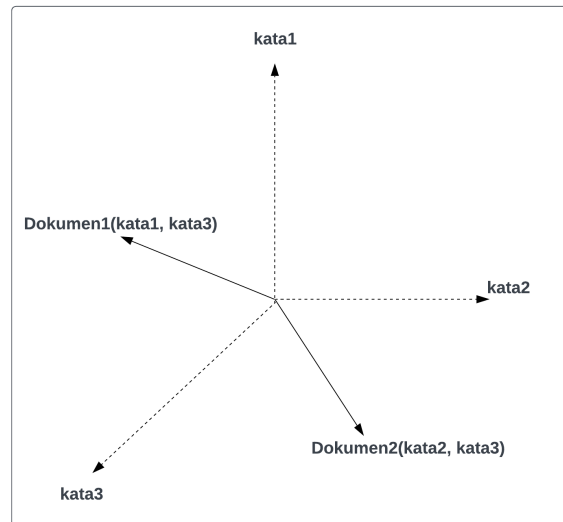
Gambar 1.1: Jumlah pengguna internet di Indonesia

Seiring dengan pesatnya perkembangan teknologi informasi, jumlah informasi yang dihasilkan khususnya data teks juga meningkat secara signifikan. Data teks ini terus dihasilkan dan berasal dari berbagai sumber seperti media sosial, situs web, *platform e-commerce* dan jurnal ilmiah *online*. Pertumbuhan yang sangat besar dari berbagai materi teks menghadirkan tantangan baru untuk pengolahan dan analisis data, terutama dalam konteks analisis data teks.

¹“Revolusi Digital: Bagaimana Komputer Mengubah Hidup Kita (Menjadi Lebih Baik atau Lebih Buruk)”, 2023, <https://medium.com/@hasonsnik/the-digital-revolution-how-computers-changed-our-lives-for-better-or-worse-d564edf6ad6c>, diakses pada 03 Maret 2024.

²“Ada 185 Juta Pengguna Internet di Indonesia pada Januari 2024”, 2023, <https://databoks.katadata.co.id/datapublish/2024/02/27/ada-185-juta-pengguna-internet-di-indonesia-pada-januari-2024>, diakses pada 03 Maret 2024.

Dalam melakukan analisis data teks seperti analisis sentimen, data teks tidak dapat langsung diolah dan dianalisis³. Oleh karena itu, umumnya data teks perlu dimodelkan ke dalam bentuk vektor, yaitu representasi teks berupa kumpulan nilai numerik. Pemodelan yang paling umum digunakan adalah *Vector Space Model* (VSM). Ide dasar VSM adalah merepresentasikan sebuah kata menjadi sebuah dimensi⁴. Jumlah dimensi mengacu pada jumlah fitur atau kata unik yang terdapat pada data teks. Pada Gambar 1.2⁵ diilustrasikan sebuah *vector space model* tiga dimensi yang dibentuk oleh tiga *terms*, yaitu “kata1”, “kata2” dan “kata3”. Pada *vector space model* tersebut terdapat dua dokumen, yaitu Dokumen1 dan Dokumen2, dimana Dokumen1 mengandung *term* “kata1” dan “kata3”, sedangkan Dokumen2 mengandung *term* “kata2” dan “kata3”. Terdapat juga istilah *Document Term Matrix* (DTM), yaitu suatu bentuk representasi data konkret yang dapat digunakan untuk menerapkan konsep VSM.



Gambar 1.2: Pemodelan data teks dalam ruang tiga dimensi

VSM dapat memodelkan data teks, namun masih memiliki kelemahan, yaitu *high dimensionality* dan *sparsity*. *High dimensionality* adalah kondisi dimana data teks memiliki jumlah fitur/kata yang sangat banyak, sedangkan *sparsity* adalah kondisi vektor berdimensi tinggi yang didominasi oleh nilai nol. *Sparsity* umum terjadi karena dari banyaknya kata yang menjadi fitur, hanya sebagian kecil yang digunakan pada suatu dokumen dalam kumpulan dokumen.

Tabel 1.1: Contoh *Document Term Matrix* (DTM)

Dokumen	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14	t15
1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0
3	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0
4	0	1	0	0	0	0	0	1	1	0	1	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1

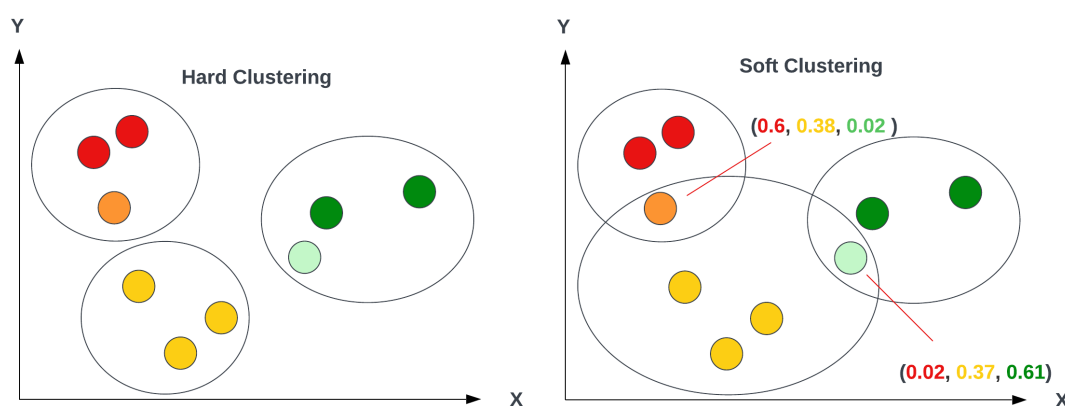
³“NLP Project Part 2: How to Clean and Prepare Data for Analysis”, 2022, <https://medium.com/@hasonsnik/the-digital-revolution-how-computers-changed-our-lives-for-better-or-worse-d564edf6ad6c>, diakses pada 03 Maret 2024.

⁴“Implementing the TF-IDF Search Engine”, 2020, https://medium.com/@kartheek_akella/implementing-the-tf-idf-search-engine-5e9a42b1d30b

⁵https://www.researchgate.net/figure/Presentation-vector-space-model_fig1_303823280

Seperti yang terlihat dalam Tabel 1.1, setiap dokumen hanya berisi tiga hingga empat kata. Namun, karena kata-kata dalam setiap dokumen berbeda, sehingga menghasilkan total 15 kata unik, maka dimensi data menjadi besar, yaitu 15 dimensi. Situasi seperti ini menyebabkan *sparsity* dan *high dimensionality*. Kedua masalah ini membuat pengolahan dan analisis data teks menjadi lebih sulit karena memerlukan waktu dan sumber daya yang lebih banyak.

Berdasarkan masalah *high dimensionality* dan *sparsity*, diusulkan teknik reduksi dimensi saat pemodelan data teks untuk menyelesaikan permasalahan. Tujuan reduksi dimensi adalah untuk mengurangi jumlah fitur tanpa menghilangkan informasi penting yang tersimpan pada fitur. Beberapa teknik reduksi dimensi yang cukup umum antara lain, *Principal Component Analysis* (PCA), *Latent Semantic Analysis* (LSA) dan *t-Distributed Stochastic Neighbor Embedding* (t-SNE). Namun pada penelitian ini dilakukan eksperimen implementasi reduksi dimensi dengan teknik klusterisasi, pendekatan *fuzzy clustering* (*soft clustering*). *Fuzzy clustering* (*soft clustering*) memungkinkan suatu objek menjadi anggota dilebih dari satu kelompok (klaster) dengan derajat keanggotaan yang berbeda di setiap kelompoknya. Pendekatan ini dapat membantu representasi yang lebih baik dari data teks yang kompleks dan memungkinkan pengelompokan yang lebih tepat dibandingkan metode *hard clustering*. Jadi pada metode *hard clustering* suatu objek hanya dapat menjadi anggota dari satu klaster walaupun sebenarnya bisa menjadi anggota dari klaster lain juga, sedangkan pada *fuzzy clustering* suatu objek memiliki persentase atau derajat keanggotaan pada setiap klaster yang ada.



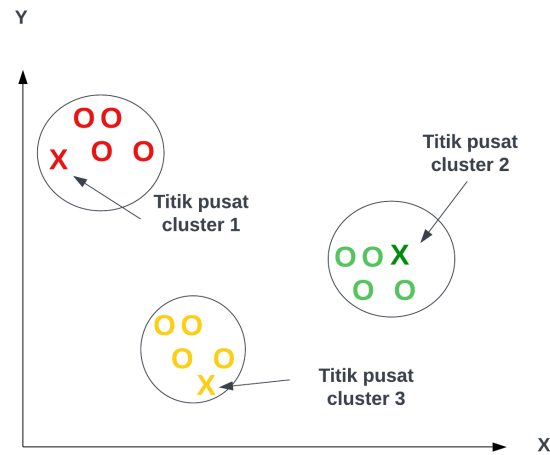
Gambar 1.3: *Hard clustering* dan *fuzzy clustering*

Perbandingan antara pengelompokan metode *hard clustering* dan *fuzzy clustering* dapat dilihat pada Gambar 1.3⁶. Diilustrasikan bahwa terdapat sembilan data yang dikelompokkan menggunakan metode *hard clustering* dan *fuzzy clustering*. Sehingga data memiliki fitur derajat keanggotaan pada setiap klaster.

Oleh karena itu penelitian ini berfokus pada implementasi reduksi dimensi dengan *soft clustering* dan *hard clustering* sebagai pembandingnya dengan alur pengerjaan yang dapat dilihat pada Gambar 1.7. Jadi dilakukan pengumpulan data terlebih dahulu dari media penyedia data. Setelah memperoleh data, data tersebut selanjutnya dibersihkan, yaitu dilakukan penghapusan karakter yang tidak memiliki arti dan dilakukan transformasi fitur yaitu mengubah kata ke bentuk kata dasar. Setelah data bersih selanjutnya adalah memodelkan data menjadi vektor, setiap dokumen dimodelkan menjadi vektor. Nilai pada setiap elemen vektor diberi bobot TF-IDF, sehingga proses ini dinamakan vektorisasi TF-IDF. Setelah data tersebut dimodelkan menjadi vektor TF-IDF, selanjutnya adalah melakukan klusterisasi. Data dikelompokkan berdasarkan kemiripan dari kata-kata yang terdapat di dalam dokumen. Dalam klusterisasi terdapat istilah yang namanya adalah

⁶https://www.researchgate.net/figure/Hard-vs-Soft-Clustering_fig1_341874469

titik pusat (*centroid*), titik pusat⁷ adalah suatu titik yang dianggap mewakili kelompok data dalam suatu kluster. Lihat pada Gambar 1.4, terdapat tiga kelompok data dengan tiga titik pusat yang dilambangkan oleh X.



Gambar 1.4: Titik pusat *cluster*

Penentuan titik pusat melibatkan inialisasi awal secara acak atau menggunakan titik-titik data tertentu. Setelah inialisasi, algoritma secara iteratif melakukan perhitungan ulang titik pusat dan pembaruan kluster sampai konvergensi tercapai. Konvergensi adalah kondisi dimana titik pusat kluster tidak mengalami perubahan yang signifikan antara iterasi sebelum dan iterasi selanjutnya. Setelah melalui proses pengelompokan, fitur dari data tersebut yang tadinya adalah kata-kata yang terkandung didalam data diubah menjadi tingkat keanggotaan antar cluster. Untuk contoh reduksi dimensi dengan *soft clustering* dimana data dikelompokkan menjadi dua *cluster*, dapat dilihat pada Gambar 3.1⁸, sebelumnya fitur dari data tersebut adalah jumlah kemunculan kata pada masing-masing dokumen sekarang berubah menjadi tingkat keanggotaan pada setiap cluster.

$$DTM = \begin{matrix} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & w_{10} \\ \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix} \rightarrow C = \begin{matrix} & c_1 & c_2 \\ \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix} & \begin{bmatrix} 0.2118281 & 0.7881719 \\ 0.8619096 & 0.1380904 \\ 0.0681949 & 0.9318051 \\ 0.8301873 & 0.1698127 \\ 0.4106981 & 0.5893019 \end{bmatrix} \end{matrix}$$

Gambar 1.5: Reduksi dimensi dengan *fuzzy clustering*

Untuk contoh reduksi dimensi dengan *hard clustering* dimana data dikelompokkan menjadi dua *cluster* dapat dilihat pada Gambar 3.2, sebelumnya fitur dari data tersebut adalah jumlah kemunculan kata pada masing-masing dokumen sekarang berubah menjadi tingkat keanggotaan pada setiap cluster. Karena implementasinya adalah *hard clustering*, maka data hanya dapat masuk kedalam satu kelompok saja sehingga penggunaan nilainya adalah biner yaitu antara angka 1 atau 0.

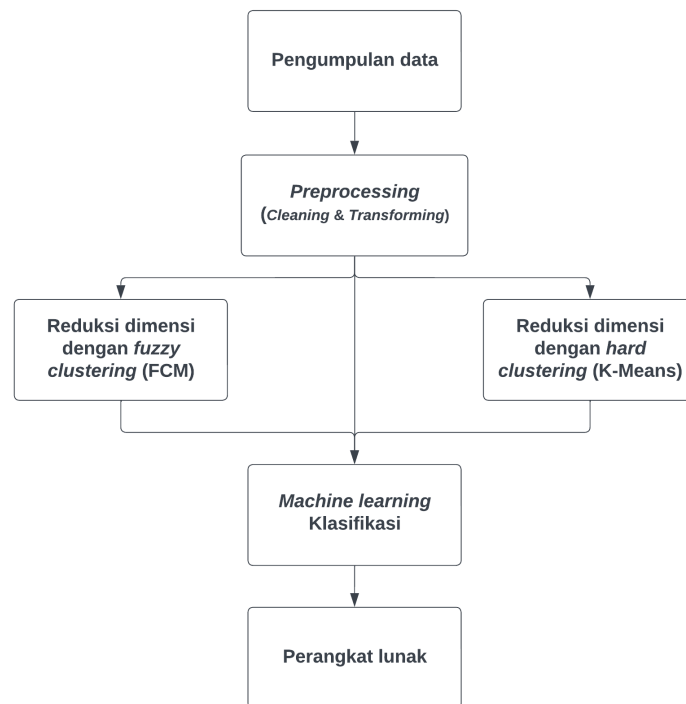
⁷<https://repositori.usu.ac.id/handle/123456789/20265>

⁸Int. J. Knowledge Engineering and Data Mining, Vol. 6, No. 3, 2019, 1–19

$$DTM = \begin{matrix} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & w_{10} \\ \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix} \rightarrow C = \begin{matrix} & c_1 & c_2 \\ \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{matrix} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{matrix}$$

Gambar 1.6: Reduksi dimensi dengan *hard clustering*

Setelah melakukan reduksi dimensi pada data, langkah selanjutnya adalah menjalankan proses klasifikasi. Untuk membandingkan metode reduksi data, penelitian ini melibatkan implementasi teknik klasifikasi dalam menyelesaikan permasalahan *task classification text* dalam konteks *sentiment analysis*. Oleh karena itu, perbandingan dilakukan antara data yang tidak mengalami reduksi dimensi dan data yang mengalami reduksi dimensi menggunakan *fuzzy clustering* dan *hard clustering*. Dua buah model klasifikasi dibangun untuk memfasilitasi perbandingan tersebut. Setelahnya, performa dari model klasifikasi dievaluasi. Akhirnya, dibuat sebuah perangkat lunak berbasis *website* yang berfungsi untuk memprediksi data baru.



Gambar 1.7: Alur proses pengerjaan TA

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dipaparkan, berikut adalah rumusan masalah yang diselesaikan dalam penelitian:

1. Bagaimana memodelkan data teks agar dimensinya tidak besar dan tidak *sparse*?
2. Bagaimana mengukur keberhasilan dari pemodelan data teks yang diusulkan?
3. Bagaimana mengimplementasikan pemodelan data teks yang diusulkan?

1.3 Tujuan

Adapun beberapa tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Memodelkan data teks menjadi derajat keanggotaan antar klaster.
2. Membandingkan performa model klasifikasi yang dilatih dan diuji menggunakan data yang fiturnya adalah derajat keanggotaan antar klaster (setelah reduksi dimensi) dan dengan model klasifikasi yang fiturnya adalah nilai TF-IDF (sebelum reduksi dimensi).
3. Membuat perangkat lunak berbasis *website* yang dapat digunakan untuk memprediksi sentimen dari data teks baru dengan mengimplementasikan pemodelan data teks usulan.

1.4 Batasan Masalah

Adapun batasan untuk penelitian ini adalah sebagai berikut:

1. Data yang digunakan adalah data teks berbahasa Inggris.

1.5 Metodologi

Metodologi yang dilakukan pada tugas akhir ini adalah sebagai berikut:

1. Melakukan studi literatur mengenai *Natural Language Processing* (NLP), teknik klasterisasi (*fuzzy clustering* dan *hard clustering*) dan teknik klasifikasi.
2. Pengumpulan data teks, dilakukan dengan mengunduh dari ai.stanford.edu. Data yang diunduh berupa *folder* yang didalamnya terdapat *file-file* ulasan film dengan format *.txt*, disimpan di dalam *folder* “neg” yang berisikan ulasan bersentimen negatif dan “pos” yang berisikan ulasan bersentimen positif.
3. Melakukan penyiapan data yaitu proses ekstrak, pembersihan dan transformasi data teks.
4. Melakukan eksplorasi data.
5. Melakukan pemodelan data menggunakan metode *Vector Space Model* (VSM).
6. Mengimplementasikan *fuzzy clustering* dan *hard clustering* untuk mereduksi dimensi data teks.
7. Mengimplementasikan klasifikasi untuk mengevaluasi efek dari reduksi dimensi.
8. Membangun perangkat lunak berbasis *website* yang dapat digunakan untuk memprediksi data ulasan baru.
9. Menulis dokumen tugas akhir.

1.6 Sistematika Pembahasan

Sistematika penulisan tugas akhir ini adalah sebagai berikut:

1. Bab 1 Pendahuluan
Bab ini berisikan penjelasan latar belakang, rumusan masalah dan tujuan dilakukan penelitian ini serta batasan masalah dan metodologi yang digunakan dalam penelitian ini.
2. Bab 2 Dasar Teori
Bab ini membahas seluruh teori yang menjadi dasar dari penelitian berdasarkan hasil dari studi literatur. Adapun teori-teori yang dibahas dalam bab ini adalah sebagai berikut:
 - (a) *Natural Language Processing*
 - (b) *Vector Space Model*
 - (c) *Document Term Matrix*
 - (d) *Regular Expression*
 - (e) *Text Preprocessing*
 - (f) *Term Weighting* TF-IDF
 - (g) *Part of Speech Tagging*
 - (h) Reduksi Dimensi

- (i) Transformasi Data melalui Normalisasi
 - (j) *Machine Learning*
 - (k) Klasterisasi
 - (l) Klasifikasi
 - (m) Visualisasi Data
3. Bab 3 Analisis Penyelesaian Masalah
Bab ini membahas mengenai deskripsi masalah yang akan diselesaikan serta tahapan dalam menyelesaikan permasalahan. Dilanjutkan dengan eksplorasi penggunaan *library* yang dimanfaatkan untuk memodelkan data teks yaitu *TfidfVectorizer* beserta dengan perhitungannya manualnya. Lalu eksplorasi penggunaan *library* yang dimanfaatkan untuk melakukan klasterisasi data yaitu *skfuzzy.cluster.cmeans* beserta dengan perhitungan manualnya.
 4. Bab 4 Penambangan Data
Bab ini membahas tentang pengumpulan, eksplorasi, dan penyiapan data ulasan yang dimanfaatkan dalam penelitian ini. Pembahasan dimulai dari deskripsi *dataset* yang diolah. Kemudian dilakukan eksplorasi terhadap *dataset* yang digunakan sehingga dapat mempermudah proses penyiapan data pada tahap selanjutnya. Lalu dilanjutkan dengan pembahasan iterasi dari eksperimen yang dilakukan beserta dengan hasil eksperimen.
 5. Bab 5 Peluncuran Model dan Pengujian
Bab ini berisi tentang perangkat lunak dalam bentuk *website* yang dapat digunakan untuk melakukan prediksi terhadap data ulasan baru, apakah termasuk kedalam sentimen negatif atau sentimen positif.
 6. Bab 6 Kesimpulan dan Saran
Bab ini berisi kesimpulan dan saran dari penelitian yang telah dilakukan.

