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 sedemikiran 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 klasterisasi, 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 dimensi nya 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 menunjukan 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

DAFTAR GAMBAR

DAFTAR TABEL

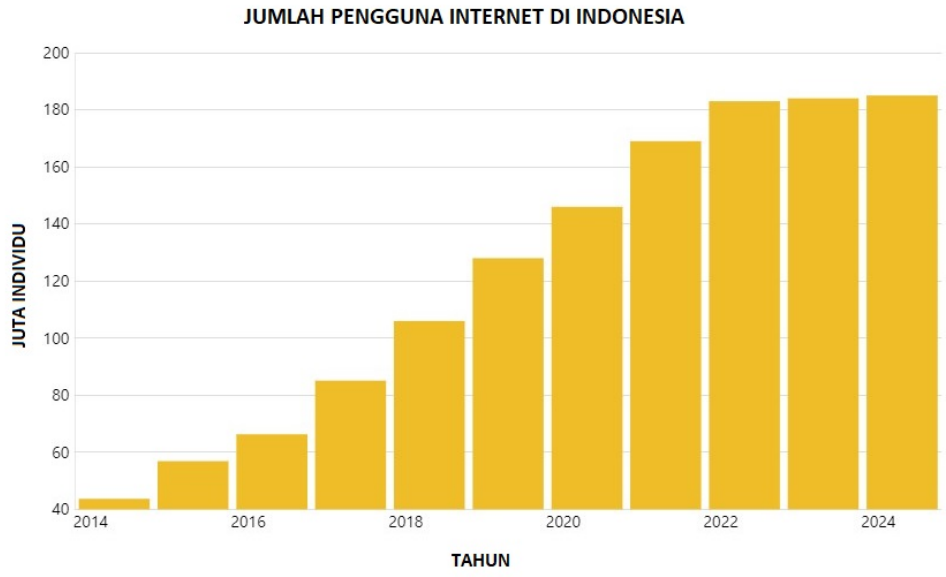
DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Teknologi komputer dan internet telah merevolusi cara manusia menjalani kehidupan[1]. 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 1.1[2]. Dari tahun 2014 sampai tahun 2024 telah terjadi penambahan pengguna internet yang sangat signifikan.



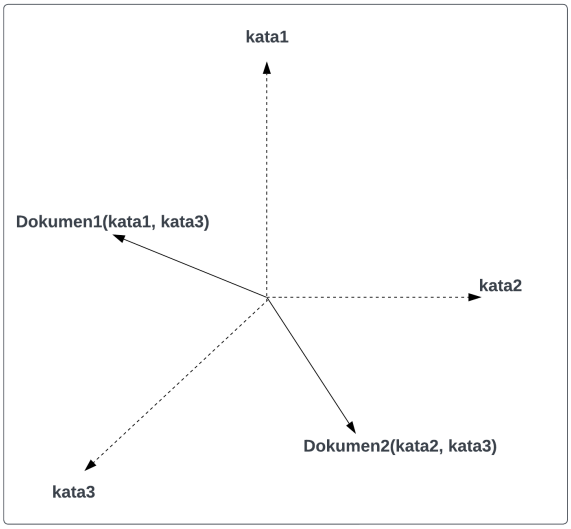
Gambar 1.1: Jumlah pengguna internet di Indonesia

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 dianalisis[3]. 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 dimensi [4]. Jumlah dimensi mengacu pada jumlah fitur atau kata unik yang terdapat pada data teks. Pada Gambar 1.2[5] 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

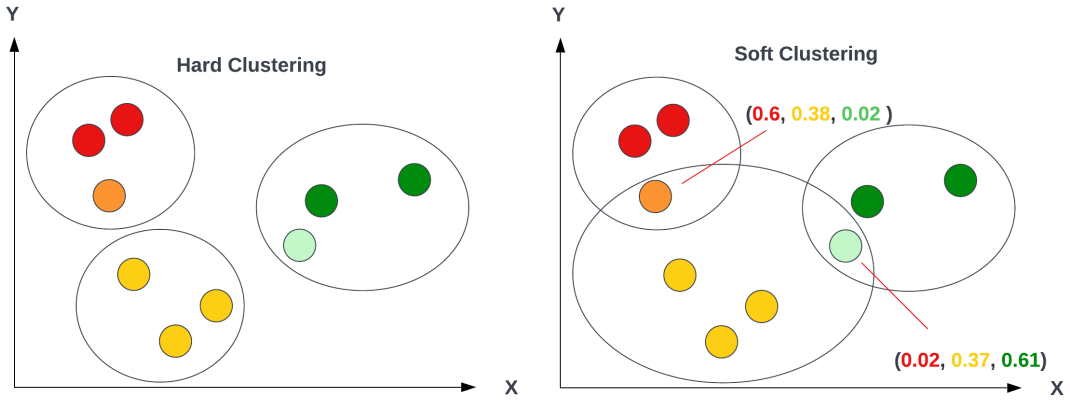
[3]"NLP Project Part 2: How to Clean and Prepare Data for Analysis", 2022, https://medium.com/@hasonsnik/the-digital-revolution-how-computers-changed-our-lives-for-better-or-worse-d564edf6ad6c, diakses pada 03 Maret 2024.

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

[5]https://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 klasterisasi, 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 pengelompokkan 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.



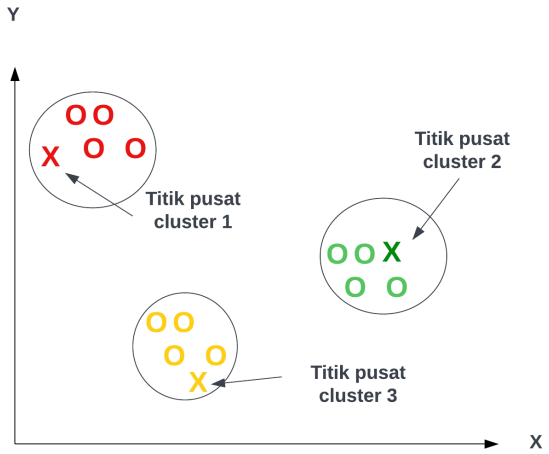
Gambar 1.3: *Hard clustering* dan *fuzzy clustering*

Perbandingan antara pengelompokkan metode *hard clustering* dan *fuzzy clustering* dapat dilihat pada Gambar 1.3[6]. Diilustrasikan bahwa terdapat sembilan data yang dikelompokkan menggunakan metode *hard clustering* dan *fuzzy clustering*. Sehingga data memiliki fitur derajat keanggotan 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 kebentuk 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 klasterisasi. Data dikelompokan berdasarkan kemiripan dari kata-kata yang terdapat di dalam dokumen. Dalam klasterisasi terdapat istilah yang namanya adalah

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

titik pusat (*centroid*), titik pusat[7] 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 inisialisasi awal secara acak atau menggunakan titik-titik data tertentu. Setelah inisialisasi, 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 pengelompokkan, 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 dikelompokan menjadi dua *cluster*, dapat dilihat pada Gambar 3.1[8], sebelumnya fitur dari data tersebut adalah jumlah kemunculan kata pada masing-masing dokumen sekarang berubah menjadi tingkat keanggotan pada setiap cluster.

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

Gambar 1.5: Reduksi dimensi dengan *fuzzy clustering*

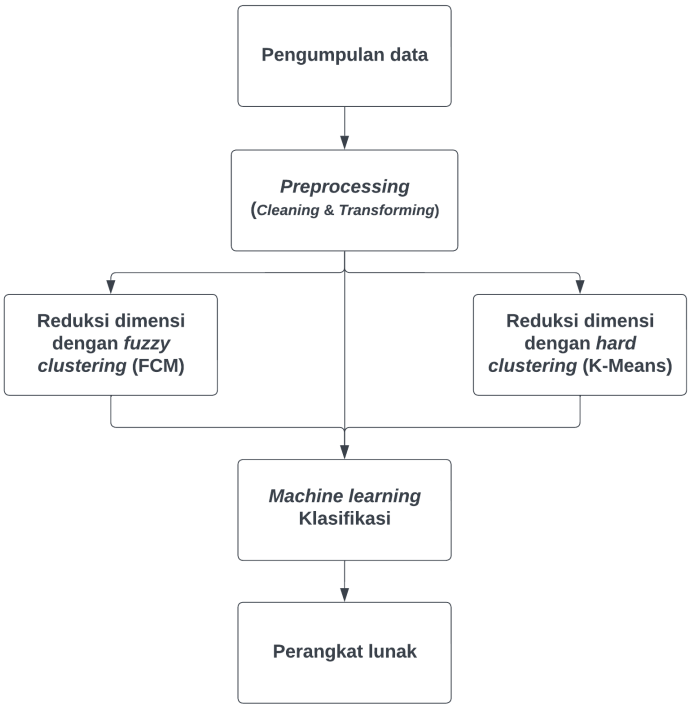
Untuk contoh reduksi dimensi dengan *hard clustering* dimana data dikelompokan menjadi dua *cluster* dapat dilihat pada Gambar 3.2, sebelumnya fitur dari data tersebut adalah jumlah kemunculan kata pada masing-masing dokumen sekarang berubah menjadi tingkat keanggotan pada setiap cluster. Karena implementasinya 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
[8] Int. J. Knowledge Engineering and Data Mining, Vol. 6, No. 3, 2019, 1–19

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

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 keanggotan 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 perhitungan 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.