Analisis Sentimen Terhadap Kebijakan Pemerintah dengan *Feature Expansion* Metode GloVe pada Media sosial Twitter

Muhammad Faiq Ardyanto Putro[1], Erwin Budi Setiawan[2]
[1,2] Universitas Telkom, Bandung
muhammadfaiq@student.telkomuniversity.ac.id[1], erwinbudisetiawan@telkomuniversity.ac.id[2]

Abstrak
Penyebaran informasi yang semakin meningkat di media sosial memudahkan pengguna untuk mengungkapkan pandangan dan pendapatnya. Opini dan reaksi dapat berupa opini positif atau negatif atau dapat diartikan sebagai sentimen. Pada riset ini dibuat sistem analisis sentimen berdasarkan data kebijakan pemerintah pada media sosial Twitter. Dalam membangun sistem analisis sentimen ini, data yang digunakan adalah data yang berisi *tweet* dengan *keyword* yang telah ditentukan dan menggunakan *feature expansion* yaitu *GloVe* dan metode klasifikasi *Support vector machine* (SVM). *feature expansion* dapat mengoreksi perbedaan kosakata dalam data *tweet* yang random dan terbatas untuk mendapatkan hasil pemrosesan kata yang maksimal. Hasil riset menunjukkan bahwa penggunaan *feature expansion GloVe* dapat meningkatkan akurasi sebesar 4,77% dari *baseline* dengan akurasi optimal sebesar 79,52%.

Kata kunci: analisis sentimen, *feature expansion*, *GloVe*, klasifikasi, SVM

Abstract
The increasing dissemination of information on social media makes it easier for users to express their views and opinions. Opinions and reactions can be positive or negative opinions or can be interpreted as sentiments. In this research, a sentiment analysis system was made based on government policy data on Twitter social media. In building this sentiment analysis system, the data used is data that contains *tweet*s with predetermined keywords and uses *feature expansion*, namely *GloVe* and the *Support vector machine* (SVM) classification method. *feature expansion* can correct for vocabulary differences in random and limited *tweet* data to get maximum word processing results. The results showed that the use of the *GloVe feature expansion* can increase the accuracy by 4.77% from the baseline with an optimal accuracy of 79.52%.

Keywords: sentiment analysis, feature expansion, GloVe, classification, SVM

1. Pendahuluan

Analisis sentimen merupakan teknik yang digunakan dalam pengekstrakan informasi yang berupa pandangan (sentimen) dari individu/kelompok terhadap isu atau kejadian yang ada. Analisis Sentimen dapat dimanfaatkan untuk mengungkap beberapa hal seperti opini publik terhadap suatu isu yang berkembang, kepuasan terhadap pelayanan, kebijakan yang dibuat, *cyber bullying*, memprediksi harga saham, serta analisis kompetitor berdasarkan data yang dikumpulkan [1].

Kebijakan pemerintah hakikatnya adalah kebijakan yang diperuntukkan kepada masyarakat baik secara langsung ataupun tidak langsung [2]. Kebijakan pemerintah dapat diartikan juga sebagai kebijakan publik dikarenakan kebijakan pemerintah disusun dan dirancang dengan sasaran tujuannya adalah publik. Menurut [3], kebijakan publik diartikan sebagai suatu kebijakan yang dapat memengaruhi setiap individu dari suatu negara. Kebijakan publik memiliki tujuan untuk mencapai kesejahteraan masyarakat melalui peraturan yang dibuat serta dapat memperoleh nilai-nilai yang diperlukan oleh warga negara untuk meningkatkan kualitas hidup baik secara fisik juga non-fisik [4].

Klasifikasi merupakan proses memprediksi kelas dari data yang diberikan. Kelas diartikan juga sebagai kategori atau label/target. Klasifikasi masuk ke dalam kategori metode *supervised learning* yang merupakan pengembangan dari ilmu *machine learning* [5]. Implementasi klasifikasi banyak dimanfaatkan di beragam disiplin ilmu, salah satunya di bidang ilmu komputer. Salah satu topik pada bidang ilmu komputer yang memanfaatkan metode klasifikasi dalam penyelesaiannya adalah Analisis sentimen. Beberapa kasus analisis sentimen yang cukup sering dibahas adalah seperti analisis sentimen mengenai suatu produk [6], tokoh publik [7], atau review terhadap suatu tempat [8]. Dalam melakukan proses klasifikasi, ada beberapa metode yang bisa digunakan untuk diimplementasikan, seperti *Support vector machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Tree, Neural Network* [9].

Deep learning adalah bagian dari kecerdasan buatan serta pembelajaran mesin dimana metode *deep learning* adalah pengembangan dari *neural network multiple layer* yang berfungsi untuk memberikan ketepatan sebagaimana dalam *object detection, speech recognition, language translation*, dan lain sebagainya. *Deep learning* saat ini banyak dikembangkan dalam berbagai sektor karena fungsi deep learning yang dapat mempermudah pekerjaan manusia terutama dalam mengolah data sehingga dapat menghasilkan informasi yang dapat dimanfaatkan [10].

Word Vector Representation (Representasi Vektor Kata) merupakan sebuah hasil pembelajaran dari

algoritma *deep learning* yang bertujuan untuk mengekstraksi kata. Terdapat beberapa metode yang dapat digunakan dalam mengimplementasikan *Word Vector Representation* diantaranya terdapat metode *Word2vec* yang telah dikembangkan oleh Google dan *GloVe* yang dikembangkan oleh *Stanford University* [11].

Berdasarkan berbagai riset dan permasalahan yang telah dijelaskan, maka pada riset ini, dilakukan analisis sentimen menggunakan *feature expansion* metode *GloVe* terhadap data kebijakan pemerintah pada media sosial twitter memanfaatkan metode klasifikasi *Support vector machine* (SVM). Alasan penggunaan teknik tersebut adalah untuk dapat mengatasi ketidakcocokan kosakata yang berasal dari data *tweet* yang kalimatnya terbatas sehingga muncul berbagai macam variasi kata, maka digunakan teknik *feature expansion* dengan metode *word embedding GloVe*.

Riset ini memiliki tujuan untuk mengetahui cara membangun sistem analisis sentimen terhadap data kebijakan pemerintah, mengetahui perbandingan hasil analisis sentimen dari penggunaan metode *GloVe* dan tanpa menggunakan *GloVe* terhadap data kebijakan pemerintah, mengetahui performansi metode klasifikasi *Support vector machine* (SVM) dalam mengklasifikasikan data kebijakan pemerintah.

Batasan yang menjadi ruang lingkup di riset ini, yaitu data yang digunakan ialah data sentimen Bahasa Indonesia yang bersumber dari twitter sebanyak 16.597 *tweet* yang berkaitan dengan kebijakan pemerintah di Indonesia, proses pelabelan sentimen dilakukan secara manual menjadi dua kategori, yaitu positif serta negatif, nilai matriks performansi yang dipakai adalah akurasi dan F1-*Score*, serta *word embedding* yang digunakan adalah *GloVe*.

Jurnal ini disusun dengan struktur sebagai berikut, pada bab 2 akan membahas teori/studi/literatur yang mendukung atau berkaitan erat dengan riset ini. Bab 3 membahas teori terkait riset dan pemodelan sistem yang dibangun. Pada bab 4 membahas hasil, analisis, dan evaluasi model riset. Lalu, pada bab 5 menjelaskan kesimpulan dan saran untuk riset selanjutnya.

2. Studi Terkait

Informasi dapat didapatkan secara masif dan mudah melalui internet dengan memanfaatkan berbagai media sosial yang ada. Peningkatan informasi yang dihasilkan dari media sosial sangatlah besar setiap harinya. Setiap terjadi suatu peristiwa atau kejadian di dunia nyata akan cepat tersebar ke seluruh dunia melalui media sosial. Setiap pengguna media sosial dapat memberikan opininya terhadap suatu peristiwa yang terjadi dan dari fenomena tersebut dapat menimbulkan suatu sentimen yang terbentuk. Sentimen yang muncul dapat berupa sentimen yang positif maupun negatif. Analisis sentimen yang terbentuk dapat menjadi acuan terhadap suatu peristiwa yang terjadi, apakah peristiwa yang terjadi tersebut merupakan suatu peristiwa yang positif atau negatif.

Terdapat berbagai riset yang telah dikembangkan pada analisis sentimen. Dalam melakukan analisis sentimen, terdapat banyak teknik *word embedding* yang bisa dipergunakan dalam melakukan analisis sentimen, seperti metode *Word2vec*, *CCA*, *GloVe*, *Doc2vec*, dan lain sebagainya. Pada riset sebelumnya mengenai analisis sentimen kolom komentar kuisioner evaluasi dosen oleh mahasiswa yang dilakukan menggunakan metode *Word2vec* menghasilkan akurasi mencapai 70%, hasil tersebut didapatkan karena data yang digunakan sedikit dan akurasi tertinggi didaptakan saat menggunakan BOW+TF-IDF dengan akurasi 85% [12].

Pada riset [13], dengan memanfaatkan *average base Word2vec* dan *bag of centroid Word2vec* dengan klasifikasi *Support vector machine* (SVM) didapatkan 85,3% saat dua model digabungkan.Pada riset lainnya, yaitu analisis sentimen calon gubernur DKI Jakarta tahun 2017 yang memanfaatkan *Lexicon Based Method* untuk menentukan *class sentiment* dan menggunakan dua teknik klasifikasi yaitu *Naïve Bayes* dan *Support vector machine* (SVM) didapatkan akurasi terbaik saat menggunakan *Naïve Bayes* membentuk akurasi hingga 95% [14].

Pada riset [15], dengan memanfaatkan metode yang sama metode klasifikasi *Support Vector Machine* (SVM) dan *Lexicon Based Features* mencapai akurasi 79%, sementara sistem analisis sentimen yang tanpa *Lexicon Based Features* mencapai akurasi yang lebih besar yaitu 84% pada parameter yang sama.

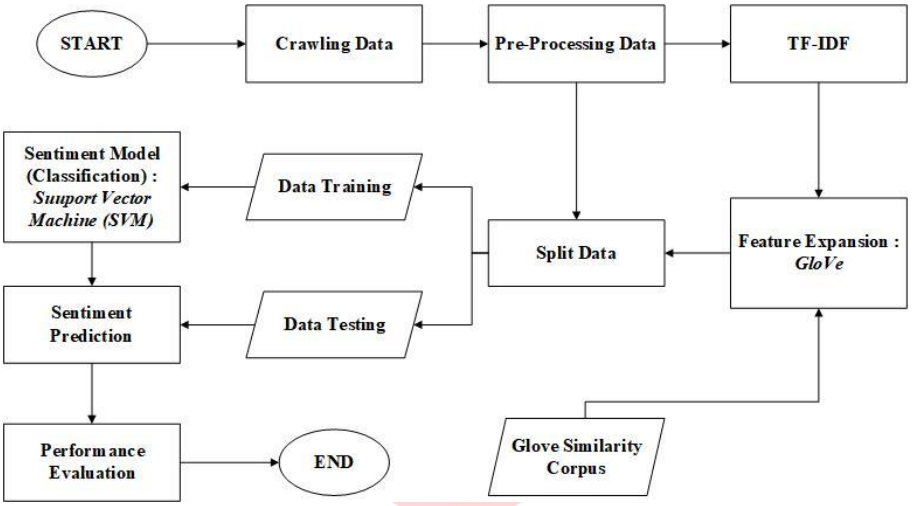
Pada riset [16], dengan memanfaatkan *feature expansion* digunakan beberapa metode *word embedding* untuk klasifikasi topik dari sebuah *tweet*. Pada riset tesebut dijelaskan bahwa panjang kalimat dari sebuah data *tweet* yang terbatas dapat mempengaruhi kalimat aslinya sehingga mengurangi makna sebenarnya dari sebuah kalimat sehingga dengan metode tersebut dapat mengurangi ketidakcocokan kosakata yang diakibatkan dari kosakata yang hilang atau kurang didalam kalimat terbatas. Pada riset ini, menunjukkan metode yang digunakan tersebut dapat meningkatkan akurasi sebesar 0,38% pada algoritma klasifikasi.

Berdasarkan berbagai riset sebelumnya dan sepengetahuan penulis, belum terdapat riset mengenai analisis sentimen menggunakan dataset dari kumpulan data *tweet* pada media sosial twitter berbahasa Indonesia yang menerapkan *feature expansion* dengan metode *GloVe*. Maka dari itu, pada penlitian ini penulis mencoba untuk membuat sebuah Sistem analisis sentimen twitter dengan *feature expansion* dan memanfaatkan metode *word embedding GloVe* dengan algoritma klasifikasi *Support Vector Machine* (SVM).

3. Sistem Analisis Sentimen Menggunakan Feature Expansion Glove

3.1 Gambaran Sistem Analisis Sentimen Menggunakan Feature Expansion Glove

Gambaran atau deskripsi sistem menggambarkan bagaimana tahap pengerjaan dalam riset. Pada riset ini membahas tentang feature expansion Word2vec untuk analisis sentimen mengenai kebijakan publik di Twitter. Alur pengerjaan pada riset ini dapat dideskripsikan pada Gambar 1.



Gambar 1. Sistem Analisis Sentimen Menggunakan Feature Expansion GloVe

3.2 *Crawling* dan Pelabelan Data

Tahapan paling awal dari pembuatan sistem analisis sentimen adalah *crawling* merupakan proses pengumpulan data baik berukuran besar maupun kecil yang berada di dalam web yang dapat disimpan di penyimpanan lokal dan data diambil sesuai dengan kata kunci yang ditentukan. Tahapan ini menggunakan Application Programming Interface (API) yang sudah dipersiapkan oleh pihak twitter. API Twitter dapat diakses dengan melakukan pengajuan otentikasi. Twitter menggunakan *Open Authentication* (OAuth) dan setiap permintaan perlu diajukan oleh pengguna Twitter yang telah resmi terdaftar. Proses crawling data pada riset ini dilakukan kurang lebih dalam 5 bulan antara bulan Maret 2021 hingga Agustus 2021.

Pada tahapan pengumpulan data akan didapatkan dataset yang berisi data campuran yang belum memiliki kelas/label. Dalam proses pembuatan dataset diperlukan proses klasfikasi. Klasifikasi merupakan proses membangun model untuk menentukan kelas atau konsep dari data. Klasifikasi mengelompokkan data menjadi beberapa kelas/label sesuai dengan yang dibutuhkan. Pada riset ini, data *tweet* yang terkumpul belum memiliki kelas/label sehingga dilakukan pelabelan secara manual (Labeling). Proses pelabelan pada dataset ini menggunakan dua kelas/label yaitu positif dan negatif. Proses pelabelan untuk tiap 1 kelas *tweet* melibatkan 3 orang lainnya dan diambil menggunakan majority vote atau suara terbanyak. Contoh dari pelabelan dataset dijelaskan di Tabel 1.

Tabel 1. Contoh Data

Tweet	Kelas/Label
alhmdulilah trimakasih gubernur terima aspirasi masyarakat khusus kaum kerja	positif
prtanyaannya sderhana produk hukum undang undang cipta kerja mnguntungkan investor buruh kaum buruh buruh protes apkah kaum buruh sgerombolan orang orang bodoh fakta serikat federasi tuai orang orang kompeten	negatif

3.3 Pre-processing Data

Data hasil *crawling* merupakan data yang masih mentah sehingga perlu dilakukan tahapan untuk membersihkan data dari informasi yang tidak diperlukan. Tahapan untuk membersihkan data yaitu *pre-processing.* Tahapan ini bertujuan agar data menjadi bersih dan dapat menghasilkan hasil yang lebih optimal ketika dilakukan pelatihan model analisis sentimen. Berikut merupakan langkah-langkah dalam *pre-processing*:

1.) *Data Cleaning*
Data cleaning adalah suatu proses yang dilakukan untuk membersihkan data yang diterapkan untuk menghilangkan noise dan memperbaiki inkonsistensi dalam data. Data dalam dunia nyata cenderung berisi noise dan tidak konsisten. Pada data cleaning *tweet*, dilakukan penghapusan URL, angka, simbol, dan atribut yang mengandung missing value atau kosong.
2.) *Stop Words*
Stop words merupakan proses yang dilakukan untuk penghapusan kata pada *tweet* yang mengandung kata yang dianggap tak berpengaruh penting dalam menentukan klasifikasi seperti konjungsi. Kata-kata tersebut dimasukkan ke dalam daftar stop words, maka kata tersebut akan dihapus dari *tweet*.
3.) *Stemming*
Stemming merupakan proses menghilangkan kata dan mengubah kata menjadi kata dasar. Proses ini dapat dilakukan dengan cara menghapus awalan atau akhiran (imbuhan) dari sebuah kata.
4.) *Case Folding*
Case folding ialah metode mengoversikan huruf kapital pada *tweet* menjadi huruf kecil. Proses ini dilakukan agar semua huruf dalam data input menjadi seragam dan mempermudah proses klasifikasi.
5.) *Tokenizing*
Tokenizing adalah sebuah tahapan memisahkan kata-kata yang dipisahkan oleh spasi. tahapan ini dilakukan untuk mempermudah klasifikasi.

3.4 *Principal Component Analysis* (PCA)

Principal Component Analysis (PCA) adalah satu teknik *feature selection* banyak digunakan. PCA mereduksi fitur yang muncul menjadi lebih sedikit tanpa menghilangkan informasi yang terdapat dalam sebuah data. Perhitungan yang digunakan pada metode PCA menggunakan nilai eigen dan vektor eigen yang menunjukkan penyebaran data dari suatu dataset. Cara kerja PCA secara sederhana yaitu fitur yang sebelumnya berjumlah n akan direduksi menjadi k, k sendiri merupakan nilai reduksi yang jumlahnya lebih kecil dari n dan dengan menggunakan k *principal component* akan nilai yang dihasilkan akan sama seperti menggunakan n [17].

3.5 *Term Frequency – Inverse Document Frequency* (TF-IDF)

Tahapan selanjutnya setelah *pre-processing*, data *tweet* kemudian ditampung dan diolah untuk diberikan nilai dari TF-IDF. Proses pembobotan dilakukan dalam setiap kalimat dalam dokumen tersebut. Nilai yang didapatkan berasal dari penjumlahan kata dari sebuah kalimat. Proses *Term Frequency* (TF) menghitung jumlah kata yang muncul dari setiap kalimat dalam dokumen tersebut, sementara *Inverse Document Frequency* (IDF) menujukkan frekuensi seberapa sering kata tersebut muncul dalam dokumen [18]. Perhitungan TF-IDF didefinisikan pada persamaan 1.

$$Wij = tfij \times Idfij$$

$$Idfij = (log\ (N/df)) \quad (1)$$

W_{ij} merupakan nilai *(weight)* dari dokmen ke-i terhadap kata ke-j, sedangkan tf_{ij} merupakan perhitungan dari kata yang dicari dalam satu dokumen. Idf_{ij} merupakan *inversed document frequency*, dimana nilai idf didapat dari nilai *log* dari total dokumen (N) dibagi dengan banyak kalimat yang mengandung kata yang dicari dalam dokumen (df).

3.6 Word Embedding GloVe

GloVe adalah teknik *word embedding* atau representasikan kata yang digunakan untuk merepresentasikan vektor dari kata dari setiap dokumen. Tujuan dari proses representasi kata menggunakan metode *GloVe* ini adalah bisa mendapatkan hubungan semantik antara kata-kata dari matriks kemunculan bersama. *Word embedding GloVe* membentuk *corpus* dengan membuat matriks kemunculan *(co-occurrence matrix)* yang menunjukkan hubungan antar kata dan seberapa sering kata tersebut muncul dalam dokumen. *GloVe* memudahkan pemrosesan pelatihan data dengan jumlah yang besar. Pada riset ini data yang dipakai ialah data *corpus tweet*. Contoh dari penerapan *word embedding GloVe* dijabarkan pada tabel 3.

Tabel 2. Contoh Dokumen

Dokumen
Aku tidak setuju dengan kebijakan ini semoga kebijakan ini ditinjau kembali

Tabel 3. GloVe

	aku	tidak	setuju	dengan	kebijakan	ini	semoga	ditinjau	kembali
aku	0	1	0	0	0	0	0	0	0
tidak	1	0	1	0	0	0	0	0	0
setuju	0	1	0	1	0	0	0	0	0
dengan	0	0	1	0	1	0	0	0	0
kebijakan	0	0	0	1	0	2	1	0	0
Ini	0	0	0	0	2	0	1	1	0
semoga	0	0	0	0	1	1	0	0	0
ditinjau	0	0	0	0	0	1	0	0	1
kembali	0	0	0	0	0	0	0	1	0

3.7 Feature Expansion

Ekspansi fitur merupakan sebuah metode yang digunakan untuk dapat mengubah nilai dari suatu fitur yang bernilai nol dengan menggunakan nilai dari kata yang similar dengan memanfaatkan *corpus similarity*. Ide dari dilakukannya ekspansi fitur adalah meningkatkan model untuk dapat meningkatkan akurasi dari algoritma klasifikasi yang digunakan. Penggunaan model ekspansi fitur dapat meningkatkan interaksi antar fitur dengan memanfaatkan *word embedding GloVe* yang menghasilkan *corpus similarity*. Proses ekspansi fitur mengubah nilai dari kata pada TF-IDF yang memiliki nilai nol dalam vektor fitur dengan persamaan kata dalam daftar *GloVe* yang muncul dalam dokumen. Ekspansi fitur ini digunakan untuk mengurangi ketidakcocokan kosakata di dalam dokumen [16].

3.8 *Support Vector Machine* (SVM)

Tahapan utama dari pembuatan sistem analisis sentimen ini adalah tahapan klasifikasi. Pada tahapan ini akan didapatkan hasil dari proses analisis sentimen yang berlangsung terhadap data yang ada. Proses klasifikasi pada riset ini memanfaatkan teknik klasifikasi *Support Vector Machine* (SVM).

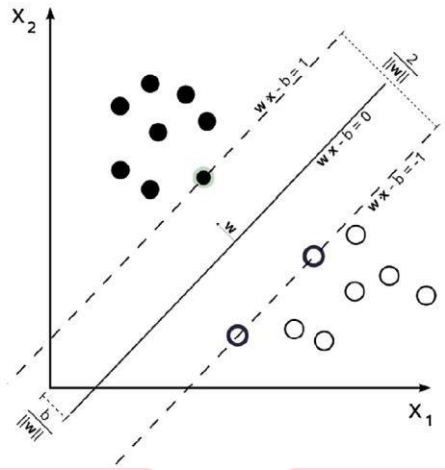
Support Vector Machine (SVM) adalah metode pada *supervised-learning* yang digunakan untuk klasifikasi. SVM memiliki konsep yang lebih matematis dibandingkan dengan metode-metode klasifikasi lainnya. Prinsip kerja SVM bekerja secara *linear* dan dikembangkan untuk dapat menyeleseaikan masalah non-*linear*. Dalam penerapannya, SVM mencari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. *Hyperplane* merupakan fungsi pemisah antar kelas [19].

Terdapat beberapa kelebihan maupun kekurangan dari metode SVM. Kelebihan SVM ialah memiliki kemampuan generalisasi data sehingga dapat menghasilkan model yang baik kendati dilatih melalui himpunan data yang sedikit hanya dengan kontrol parameter yang simpel serta relatif mudah dalam pengimplementasiannya. Sementara kelemahan dari metode SVM akan cukup sulit untuk diaplikasikan dalam kelompok data dengan spesimen dan dimensi besar. Penjelasan mengenai Metode SVM dapat dilihat dari definisi pada persamaan 2.

$$\{x_i, y_i\} \;\; dimana \;\; i = 1 \ldots n, \qquad y_i \in \{-1, 1\}, \qquad x \in \mathbb{R}^D$$
$$\vec{w} \,.\, x_i + b = 0$$
$$\vec{w} \,.\, x_i + b \leq -1, \qquad untuk \; y_i = -1$$
$$\vec{w} \,.\, x_i + b \geq 1, \qquad untuk \; y_i = 1 \quad (2)$$

Dari definisi diatas, diberikan data pelatihan sejumlah n, dengan input x, dengan D merupakan dimensi vektor dan y_i bernilai 1 atau -1. *Hyperplane* yang baik adalah yang dapat memaksimalkan jarak antar *margin* (antara 1 sampai -1). Lalu, jika mempertimbangkan titik-titik yang posisinya paling dekat (*support vector*) dengan garis pemisa *hyperplane*, maka dua bidang H_1 dan H_2 tempat posisi dari titik-titik tersebut didefinisikan seperti pada persamaan 3.

$$\vec{w} \,.\, x_i + b = +1, \qquad untuk \; H_1$$
$$\vec{w} \,.\, x_i + b = -1, \qquad untuk \; H_2 \quad (3)$$



Gambar 2. Support Vector Machine

3.9 Confusion Matrix

Confusion Matrix merupakan salah satu teknik yang bisa dipergunakan dalam mengukur kinerja dari suatu metode klasifikasi. *Confusion Matrix* berisi informasi tentang klasifikasi aktual dan prediksi yang dilakukan oleh sistem klasifikasi.

Confusion Matrix digunakan untuk menganalisis seberapa baik sistem klasifikasi yang telah dibuat dalam mengenali tuple dari kelas yang berbeda. Pada pengukuran menggunakan *Confusion Matrix*, istilah yang digunakan dalam *Confusion Matrix* antara lain *True Positive* (TP) yaitu data positif yang terklasifikasikan secara akurat, *True Negative* (TN) ialah data negatif yang terklasifikasikan secara akurat, *False Positive* (FP) ialah data negatif yang terklasifikasikan tidak akurat, dan *False Negative* (FN) data positif yang terklasifikasikan tidak akurat [20]. Tabel pembagian *Confusion Matrix* dijelaskan pada tabel 4.

Tabel 4. Confusion Matrix

Kelas	Prediksi	
	Positif	Negatif
Positif	*True Positive* (TP)	*False Negative* (FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)

Dari tabel *Confusion Matrix* diatas diperoleh nilai akurasi, presisi, dan *recall*. Akurasi merupakan proporsi dari jumlah total prediksi yang akurat. Akurasi merefleksikan ketepatan sistem dalam mengklasifikasikan data. Akurasi ditentukan dengan persamaan:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2)$$

Presisi (PRE) menunjukkan total data berlabel positif yang diklasifikasikan secara tepat dibagi dengan total data yang diklasifikasi positif. Nilai PRE ditentukan dengan persamaan:

$$\text{PRE} = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

Recall (RE) menyatakan berapa persentase data berlabel positif yang diklasifikasikan dengan tepat oleh sistem. Nilai RE ditentukan dengan persamaan:

$$\text{RE} = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

F1-*Score* ialah *harmonic mean* dari *recall* serta presisi. Nilai rata-rata harmonik dihitung dengan mengubah nilai presisi dan *recall* ke bentuk pecahan. F1-*Score* ditentukan dengan persamaan:

$$\text{F1} - \text{Score} = 2 \times \frac{PRE \times RE}{PRE + RE} \quad (5)$$

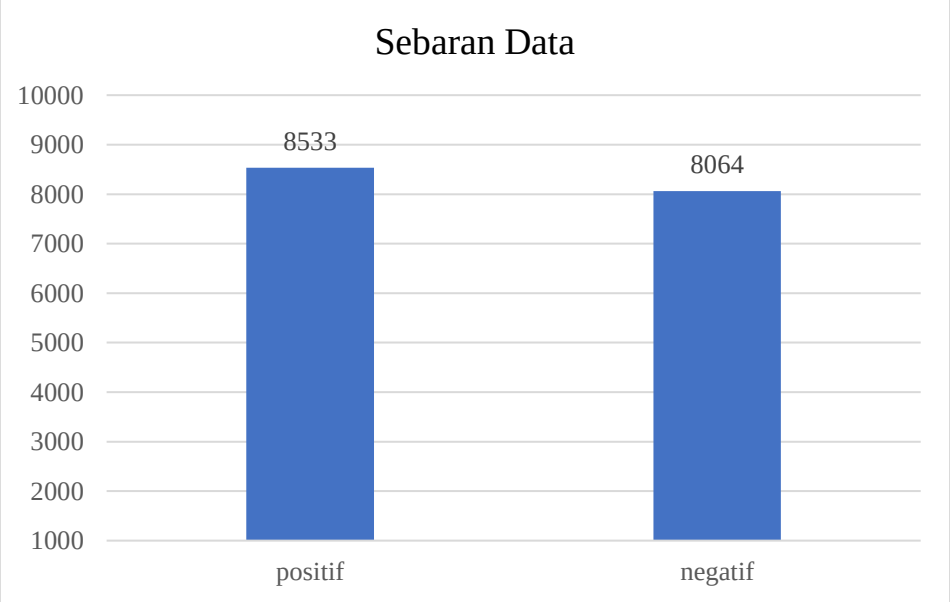
4. Evaluasi

4.1 Data

Data *tweet* yang digunakan sebanyak 16.597 *tweet* berbahasa Indonesia dengan topik kebijakan pemerintah di Indonesia yang terdiri dari beberapa *keyword* yaitu #uuite, #omnibuslaw, #ruukpk, #uuciptakerja, #ppkm, #reformasidikorupsi #dewanperampokrakyat, dan #mositidakpercaya. Data dibagi menjadi data berlabel positif dan negatif. Persebaran data yang sudah dilabeli dapat dilihat pada tabel 5 dan gambar 3.

Tabel 5. Persebaran Data dalam persen

Sentimen	Persebaran
Positif	51,41%
Negatif	48,59%



Gambar 3. Persebaran Data

4.2 Pre-processing

Berikut ini adalah contoh dari data *tweet* yang telah dilakukan tahapan *pre-processing*

1. *Lowercase*

No	Sebelum	Sesudah
1.	Bilangnya begini, maksudnya begitu ucap Pak Sapardi, Yo Ndak Tau Kok Tanya Saya tiba-tiba saut Pak Jokowi. Ya kesekian kalinya pemimpin satu ini berucap akan tetapi selalu saja paradoks dengan realitanya. Dari penolakan UU Revisi KPK, Revisi UU Minerba hingga UU Ciptaker, https://t.co/slKzkDzqiv	**b**ilangnya begini, maksudnya begitu ucap **p**ak **s**apardi, **y**o **n**dak **t**au **k**ok **t**anya **s**aya tiba-tiba saut **p**ak **j**okowi. **y**a kesekian kalinya pemimpin satu ini berucap akan tetapi selalu saja paradoks dengan realitanya. dari penolakan **uu** **r**evisi **kpk**, **r**evisi **uu** **m**inerba hingga **uu** **c**iptaker, https://t.co/slkzkdzqiv

2. *Special Text Removal (Cleaning)*

No	Sebelum	Sesudah

1.	bilangnya begini**,** maksudnya begitu ucap pak sapardi**,** yo ndak tau kok tanya saya tiba-tiba saut pak jokowi**.** ya kesekian kalinya pemimpin satu ini berucap akan tetapi selalu saja paradoks dengan realitanya**.** dari penolakan uu revisi kpk**,** revisi uu minerba hingga uu ciptaker**, https://t.co/slkzkdzqiv**	bilangnya begini maksudnya begitu ucap pak sapardi yo ndak tau kok tanya saya tiba tiba saut pak jokowi ya kesekian kalinya pemimpin satu ini berucap akan tetapi selalu saja paradoks dengan realitanya dari penolakan uu revisi kpk revisi uu minerba hingga uu ciptaker

3. Normalisasi Kata

No	Sebelum	Sesudah
1.	bilangnya begini maksudnya begitu ucap pak sapardi yo ndak tau kok tanya saya tiba tiba saut pak jokowi ya kesekian kalinya pemimpin satu ini berucap akan tetapi selalu saja paradoks dengan realitanya dari penolakan **uu** revisi **kpk** revisi **uu** minerba hingga **uu** ciptaker	bilangnya begini maksudnya begitu ucap pak sapardi yo ndak tau kok tanya saya tiba tiba saut pak jokowi ya kesekian kalinya pemimpin satu ini berucap akan tetapi selalu saja paradoks dengan realitanya dari penolakan **undang undang** revisi **komisi pemberantasan korupsi** revisi **undang undang** minerba hingga **undang undang** ciptaker

4. *Stopwords*

No	Sebelum	Sesudah
1.	bilangnya begini maksudnya **begitu** ucap pak sapardi yo **ndak** tau kok tanya **saya** tiba tiba saut pak jokowi ya kesekian kalinya pemimpin satu **ini** berucap **akan** tetapi selalu **saja** paradoks **dengan** realitanya **dari** penolakan undang undang revisi komisi pemberantasan korupsi revisi undang undang minerba hingga undang undang ciptaker	bilangnya begini maksudnya ucap pak sapardi yo tau kok tanya tiba tiba saut pak jokowi ya kesekian kalinya pemimpin satu berucap tetapi selalu paradoks realitanya penolakan undang undang revisi komisi pemberantasan korupsi revisi undang undang minerba hingga undang undang ciptaker

5. *Stemming*

No	Sebelum	Sesudah
1.	bilangnya begini maksudnya ucap pak sapardi yo tau kok tanya tiba tiba saut pak jokowi ya kesekian kalinya pemimpin satu berucap tetapi selalu paradoks realitanya penolakan undang undang revisi komisi pemberantasan korupsi revisi undang undang minerba hingga undang undang ciptaker	**bilang** begini **maksud** ucap pak sapardi **iya** tau kok tanya tiba tiba saut pak jokowi **iya sekian kali pimpin** satu **ucap** tetapi selalu paradoks **realitas tolak** undang undang revisi komisi **berantas** korupsi revisi undang undang **mineral batubara** hingga undang undang **cipta kerja**

6. *Tokenizing*

No	Sebelum	Sesudah
1.	bilang begini maksud ucap pak sapardi iya tau kok tanya **tiba tiba** saut pak jokowi iya sekian kali pimpin satu ucap tetapi selalu paradoks realitas tolak **undang undang** revisi **komisi berantas korupsi** revisi **undang undang mineral batubara** hingga **undang undang cipta kerja**	['bilang', 'begini', 'maksud', 'ucap', 'pak', 'sapardi', 'iya', 'tau', 'kok', 'tanya', 'tiba', 'tiba', 'saut', 'pak', 'jokowi', 'iya', 'sekian', 'kali', 'pimpin', 'satu', 'ucap', 'tetapi', 'selalu', 'paradoks', 'realitas', 'tolak', 'undang', 'undang', 'revisi', 'komisi', 'berantas', 'korupsi', 'revisi', 'undang', 'undang', 'mineral', 'batubara', 'hingga', 'undang', 'undang', 'cipta', 'kerja']

4.3 Pembuatan *Corpus*

Pada riset ini, pembuatan kamus kata atau *corpus* digunakan metode *word embedding GloVe* dengan parameter *window* dengan ukuran 3. *Corpus* yang dibuat yang berasal dari data tweet yang telah dikumpulan sebelumnya, *Corpus* ini berisi kumpulan kata yang telah diurutkan nilai similaritasnya. Dari data tweet sebanyak 16597 didapatkan sebanyak 10869 kosakata.

Tabel 6. Corpus Similarity

Kata	**Peringkat 1**	**Peringkat 2**	**Peringkat 3**	**Peringkat 4**	**Peringkat 5**
Covid	vaksin	pandemi	sehat	biar	nih
	Peringkat 6	**Peringkat 7**	**Peringkat 8**	**Peringkat 9**	**Peringkat 10**
	dapat	terima	tangan	vaksinasi	kasih

4.4 Pembobotan TF-IDF

Setelah data selesai dilakukan proses *pre-processing* tahapan selanjutnya adalah melakukan pembobotan kata dari setiap kalimat pada data tersebut dengan metode TF-IDF. Proses ini untuk mengetahui nilai kemunculan kata pada dokumen tersebut.

4.5 Feature Expansion

Data yang telah melalui proses *pre-processing* dan pembobotan kata TF-IDF akan melalui proses *feature expansion* dimana pada proses ini kata yang memiliki ketidakcocokan atau memiliki nilai nol akan diganti dengan kata yang *similar* sesuai dengan corpus similarity yang telah dibuat. Sebagai contoh pada pembobotan kata TF-IDF kata “covid” memiliki nilai nol. Namun, pada salah satu dokumen *tweet* tersebut berisi kalimat “vaksin biar caem undang undang cipta kerja beri mudah bangun usaha”. Lalu dilakukan pengecekan pada corpus similarity yang telah dibuat sebelumnya kata “covid” memiliki kata similaritas dengan kata “vaksin” maka nilai nol dari kata “covid” diganti dengan nilai dari kata “vaksin”.

Tabel 7. Representasi Vektor Dokumen 1

Dokumen	vaksin	biar	caem	undang	undang	…	kerja	…	**covid**
1	1	1	1	1	1	…	1	…	**0**

Sebagai contoh pada tabel 7 pembobotan kata TF-IDF kata “covid” memiliki nilai nol. Namun, pada salah satu dokumen *tweet* tersebut berisi kalimat “vaksin biar caem undang undang cipta kerja beri mudah bangun usaha”. Lalu dilakukan pengecekan pada corpus similarity yang telah dibuat sebelumnya kata “covid” memiliki kata similaritas dengan kata “vaksin” maka nilai nol dari kata “covid” diganti dengan nilai dari kata “vaksin”.

Tabel 8. Representasi Vektor Dokumen 1 setelah Ekspansi Fitur

Dokumen	vaksin	biar	caem	undang	undang	…	kerja	…	**covid**
1	1	1	1	1	1	…	1	…	**1**

Proses ekspansi fitur ini dilakukan untuk keseluruhan kata yang terdapat pada data *tweet* yang sudah melewati tahapan pembobotan dengan memanfaatkan *corpus* yang telah dibuat.

4.6 Skenario dan Hasil Pengujian

Pada riset ini, digunakan 3 skenario pengujian dengan algoritma klasifikasi *Support Vector Machine.* Skenario yang pertama adalah melakukan pengujian dengan memanfaatkan metode *feature selection Principal Component Analysis (PCA)* yang digunakan sebagai *baseline*. Skenario yang kedua adalah melakukan pengujian dengan memasukkan proses pembobotan TF-IDF untuk mengetahui pengaruh dengan adanya pembobotan TF-IDF. Skenario yang ketiga adalah dengan menambahkan *feature expansion* untuk mengetahui pengaruh dari adanya penambahan proses *feature expansion*.

Tabel 9. Hasil Performansi Baseline

Rasio	Akurasi (%)	F1-Score
70:30	75,78%	0,7584
80:20	75,42%	0,7524
90:10	**75,90%**	**0,7561**

Berdasarkan hasil yang didapatkan pada tabel 9, didapatkan hasil paling optimal pada rasio pembagian data latih dan data uji dengan rasio 90:10 dengan nilai akurasi sebesar 75,90%, sehingga untuk pengujian selanjutnya akan menggunakan rasio pembagian data latih dan data uji 90:10.

Tabel 10. Hasil Performansi Baseline + TF-IDF

Model	Akurasi (%)	F1-Score
Baseline	75,90%	0,7561
Baseline + TF-IDF	**78,37% (+3,25)**	**0,7818 (+3,40)**

Pada skenario kedua didapatkan hasil seperti pada tabel 10, pembobotan TF-IDF meningkatkan akurasi dari algoritma klasifikasi sebesar 3,25% dari *baseline* menjadi 78,37%, serta meningkatkan nilai *f1-score* sebesar 3,40% menjadi 0,7818.

Tabel 11. Hasil Performansi Baseline + TF-IDF + GloVe (TOP 1)

Random State	Akurasi (%)	F1-Score
1	78,37% (+3,25)	0,7818 (+3,40)
24	77,95% (+2,70)	0,7779 (+2,88)
38	78,73% (+3,73)	0,7867 (+4,05)
42	79,40% (+4,61)	0,7927 (+4,84)
54	**79,52% (+4,77)**	**0,7942 (+5,04)**

Pada skenario ketiga dengan fitur TOP 1 didapatkan hasil akurasi yang meningkat hingga 4,77% dari 75,90% menjadi 79,52%, serta meningkatkan nilai *f1-score* sebesar 5,04% menjadi 0,7942.

Tabel 12. Hasil Performansi Baseline + TF-IDF + GloVe (TOP 5)

Random State	Akurasi (%)	F1-Score
1	78,19% (+3,02)	0,7803 (+3,20)

24	77,89% (+2,62)	0,7769 (+2,75)
38	79,16% (+4,30)	0,7908 (+4,59)
42	79,34% (+4,53)	0,7902 (+4,51)
54	**79,34% (+4,53)**	**0,7920 (+4,75)**

Pada skenario ketiga dengan fitur TOP 5 didapatkan hasil akurasi yang meningkat hingga 4,53% dari *baseline* menjadi 79,52%, serta meningkatkan nilai *f1-score* sebesar 4,75% menjadi 0,7920.

Tabel 13. Hasil Performansi Baseline + TF-IDF + GloVe (TOP 10)

Random State	Akurasi (%)	F1-Score
1	78,25% (+3,10)	0,7811 (+3,31)
24	78,19% (+3,02)	0,7798 (+3,13)
38	78,67% (+3,65)	0,7862 (+3,98)
42	**79,34% (+4,53)**	**0,79 (+4,48)**
54	79,04% (+4,14)	0,7886 (+4,30)

Pada skenario ketiga dengan fitur TOP 10 didapatkan hasil akurasi yang meningkat hingga 4,53% dari Baseline menjadi 79,34%, serta meningkatkan nilai *f1-score* sebesar 4,48% menjadi 0,79.

4.7 Analisis Hasil Pengujian

Berdasarkan dari hasil pengujian diatas, pengujian meggunakan *feature expansion* dapat menghasilkan hasil yang berbeda-beda tergantung daripada penggunaan *corpus* kata dan ukuran fitur yang digunakan. Dari pengujian yang telah dilakukan, proses pembobotan TF-IDF dan feature expansion dapat meningkatkan akurasi dengan cukup baik dibandingkan dari tanpa menggunakan pembobotan TF-IDF dan *feature expansion*. Nilai optimum akurasi yang didapakan dari pengujian yang dilakukan adalah 79,52% serta nilai f1-score mencapai 0,7942 pada fitur top 1 dan random state 54. Hal tersebut dikarenakan fitur top 1 akan lebih spesifik dalam mencari similaritas dari kata.

5. Kesimpulan

Pada riset ini, dilakukan analisis sentimen menggunakan *feature expansion GloVe* dan algoritma klasifikasi *Support Vector Machine*. Pada riset ini, diketahui bahwa dalam membangun sistem analisis sentimen terhadap data kebijakan pemerintah dibutuhkan beberapa tahap mulai dari *crawling* data, *pre-processing*, pembobotan TF-IDF, pembuatan *corpus*, *feature expansion GloVe*, serta proses klasifikasi. Berdasarkan hasil riset diatas juga dapat disimpulkan bahwa penggunaan *feature expansion GloVe* terbukti dapat meningkatkan akurasi dari algoritma klasifikasi dibandingkan dengan tanpa menggunakan *feature expansion* dengan peningkatan akurasi sebesar 4,77%. Dari pengujian diatas didapatkan juga hasil paling optimal untuk akurasi pada algoritma klasifikasi Support Vector Machine pada fitur Top 1 dengan nilai akurasi sebesar 79,52% dan nilai F1-Score sebesar 0,7942.

REFERENSI

[1] I. Septiar, “Introduction: Sentiment Analysis. Mudah?,” *medium.com*, 2019. https://medium.com/@irvanseptiar/introduction-sentiment-analysis-mudah-5785f88e435d (accessed

Oct. 22, 2020).

[2] P. KASN, "Studi Kebijakan Pemerintah," *perpustakaan.kasn.go.id/*. http://perpustakaan.kasn.go.id/index.php?p=show_detail&id=175&keywords=#:~:text=Kebijakan pemerintah pada hakikatnya merupakan,pada berbagai dimensi kehidupan publik. (accessed Jan. 19, 2021).

[3] Cambridge, "Public Policy," *dictionary.cambridge.org/*. https://dictionary.cambridge.org/dictionary/english/public-policy (accessed Jan. 19, 2021).

[4] A. S. Putri, "Kebijakan Publik: Pengertian, Tujuan dan Ciri-ciri," *kompas.com*, 2020. https://www.kompas.com/skola/read/2020/02/06/210000269/kebijakan-publik--pengertian-tujuan-dan-ciri-ciri?page=all (accessed Dec. 19, 2021).

[5] S. Asiri, "Machine Learning Classifiers," *towardsdatascience.com*, 2018. https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623#:~:text=Classification is the process of,discrete output variables (y). (accessed Nov. 19, 2020).

[6] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[7] A. Fathan Hidayatullah and A. Sn, "ISSN: 1979-2328 UPN "Veteran," *Semin. Nas. Inform.*, vol. 2014, no. semnasIF, pp. 115–122, 2014, [Online]. Available: http://www.situs.com.

[8] D. A. Muthia, "Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes," *Jurnalilmu Pengetah. Dan Teknol. Komput.*, vol. 2, no. 2, pp. 39–45, 2017.

[9] S. Dewi, "Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan," *Techno Nusa Mandiri*, vol. XIII, no. 1, pp. 60–66, 2016.

[10] Iykra, "Deep Learning dan Manfaatnya Bagi Perkembangan AI," *medium.com*, 2016. https://medium.com/iykra/deep-learning-dan-manfaatnya-bagi-perkembangan-ai-cab94e20c19a (accessed Nov. 19, 2020).

[11] D. Suhartono, "Word Vector Representation: Word2Vec & Glove," *socs.binus.ac.id*, 2016. https://socs.binus.ac.id/2016/12/22/word-vector-representation-word2vec-glove/#:~:text=Word2Vec merupakan nama word vector,mengukur beberapa vektor sebagai perbandingan. (accessed Nov. 19, 2020).

[12] M. A. Fauzi, "Word2Vec model for sentiment analysis of product reviews in Indonesian language," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, p. 525, 2019, doi: 10.11591/ijece.v9i1.pp525-530.

[13] M. Rusli, "Ekstraksi Fitur Menggunakan Model Word2Vec Pada Sentiment Analysis Kolom Komentar Kuisioner Evaluasi Dosen Oleh Mahasiswa," *Klik - Kumpul. J. Ilmu Komput.*, vol. 7, no. 1, p. 35, 2020, doi: 10.20527/klik.v7i1.296.

[14] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *INTEGER J. Inf. Technol.*, vol. 1, no. 1, pp. 32–41, 2017, [Online]. Available: https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf.

[15] U. Rofiqoh, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 12, pp. 1725–1732, 2017, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/628.

[16] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic classification," *Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol.*, no. October, 2017, doi: 10.1109/TSSA.2016.7871085.

[17] N. A. S. Galih Hendro M, T.B. Adji, "Penggunaan Metodologi Analisa Komponen Utama (PCA) untuk Mereduksi Faktor-Faktor yang Mempengaruhi Penyakit Jantung Koroner.pdf." 2012.

[18] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[19] Samsudiney, "Penjelasan Sederhana tentang Apa Itu SVM?," *medium.com*, 2019.

https://medium.com/@samsudiney/penjelasan-sederhana-tentang-apa-itu-svm-149fec72bd02 (accessed Nov. 23, 2020).

[20] J. Han, M. Kamber, and J. Pei, "Data Mining Techniques, Third Edition," p. 847, 2011.

Lampiran

Lampiran dapat berupa detil data dan contoh lebih lengkapnya, data-data pendukung, detail hasil pengujian, analisis hasil pengujian, detail hasil survey, surat pernyataan dari tempat studi kasus, screenshot tampilan sistem,hasil kuesioner dan lain-lain.