METODE *FASTICA* UNTUK REDUKSI DATA DIMENSI TINGGI PADA ANALISIS SENTIMEN PARIWISATA KOTA SEMARANG MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

Mochamad Amry Assiva[1], Heru Agus Santoso[2], Catur Supriyanto[3]
[123]Pascasarjana Teknik Informatika Universitas Dian Nuswantoro

Abstract

Some communities have a voice attractions via Twitter. The opinion can be used as sentiment analysis to determine the ratings of a tourist attraction. Results of sentiment analysis is expected to assist in the improvement and evaluation of the attraction. In related research sentiment analysis previously used linear dimension reduction method, but has the disadvantage produce a linear combination of all the features that will have difficulty if dealing with data that is non-linear. Therefore, in this study used methods of non-linear dimension reduction, namely FastICA in order to improve the accuracy of Support Vector Machine classifier that can handle high-dimensional and non-linear data. This study uses the Indonesian language text contained on the social networking site Twitter. Validation is done by using a 10-Fold Cross Validation. While the measurement accuracy is measured by the Confusion Matrix and ROC curves. Results application of dimension reduction FastICA gain accuracy of 92.90% and the AUC 0.9157 which means the accuracy of 0.95% better than on Support Vector Machine itself, is proven to increase the accuracy of the SVM algorithm on the non-linier tweet data of attractions in the city of Semarang that can be classified by both in positive and negative class..

Keywords: Sentiment Analysis, SVM, FastICA

1. PENDAHULUAN

Sebagai daerah tujuan wisata utama di Pulau Jawa, Kota Semarang mempunyai berbagai macam wisata yang menarik. Kebanyakan wisatawan datang ke Kota Semarang untuk melihat kebudayaan dan tradisi Jawa yang masih kuat. Pariwisata bagi pemerintah daerah merupakan salah satu aspek untuk meningkatkan pendapatan daerah. Salah satu kendala yang dihadapi oleh pemerintah daerah dalam hal pengembangan pariwisata adalah kurang baiknya pengelolaan fasilitas objek wisata. Sektor wisata yang beragam dengan keunikannya dan didukung dengan fasilitas serta sarana transportasi yang tersedia di kawasan wisata dapat memberikan pendapatan pemerintah yang sangat besar.

Salah satu media yang efektif untuk menampung opini objek wisata ini adalah *Twitter*, yang termasuk cepat dalam memberikan informasi tentang pengalaman yang dirasakan oleh masyarakat sebagai bahan evaluasi untuk pihak yang mengelola objek wisata. Selain itu *Twitter* sendiri merupakan salah satu media sosial yang akrab digunakan oleh masyarakat Indonesia, yang tentunya akan memudahkan untuk pengumpulan opini dibandingkan dengan melakukan survey ataupun penyebaran kuisioner.

Salah satu cabang riset yang kemudian berkembang dari situasi ledakan informasi di internet adalah *sentiment analysis*. Sentimen Analisis atau *opinion mining* adalah studi komputasional dari opini-opini orang, *appraisal* dan emosi melalui entitas, *event* dan atribut yang dimiliki [1]. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek, apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif atau negatif.

Salah satu kesulitan klasifikasi sentimen teks adalah tingginya dimensi fitur yang digunakan untuk teks, yang menimbulkan rintangan yang besar dalam menerapkan banyak algoritma *machine learning* untuk klasifikasi sentimen teks. Salah satu teknik untuk mengatasi tingginya dimensi fitur adalah seleksi fitur, berbagai teknik seleksi fitur [2], *Information Gain* [3], telah digunakan untuk mengurangi dimensi

vektor. Tujuan dari metode seleksi fitur adalah untuk mendapatkan pengurangan set dengan menghapus beberapa fitur yang dianggap fitur tidak relevan untuk klasifikasi sentimen teks agar menghasilkan peningkatan klasifikasi akurasi dan penurunan durasi algoritma *machine learning* [4]. Namun kelemahan seleksi atribut yaitu memerlukan pelatihan satu set data besar untuk mendapatkan transformasi yang dapat diandalkan [5].

Selain teknik seleksi fitur, teknik lain yang dapat digunakan untuk mengatasi masalah tingginya dimensi fitur adalah reduksi dimensi. Tujuan dari teknik pengurangan dimensi adalah untuk mendapatkan representasi data baru yang dikelola menjadi dimensi lebih rendah [6]. Ekstraksi atribut secara umum diklasifikasikan menjadi linier dan nonlinier [7]. Algoritma linier reduksi dimensi terdiri dari algoritma *Singular Value Decomposition* (SVD) dan algoritma *Principal Component Analysis* (PCA) [8]. Namun kelemahan algoritma linier reduksi dimensi adalah menghasilkan kombinasi linier dari semua fitur yang mungkin akan terkontaminasi oleh noise dan menurunkan kinerja algoritma klasifikasi, algoritma linier reduksi dimensi akan menemui kesulitan jika berhadapan dengan data-data yang non-linier [9].

Salah satu metode non-linier reduksi dimensi adalah *Independent Component Analysis* (ICA). ICA dapat menemukan representasi linier dari data *non-gaussian* sehingga komponen dari data tersebut independen secara statistik, atau se-independen mungkin, representasi tersebut mampu menangkap struktur penting dari data. ICA memiliki karakteristik proses yang berbeda, seperti non-linier, dinamis atau *multi-modality* [6].

Pengembangan dari algoritma *Independent Component Analysis* adalah *FastICA*, sebuah algoritma yang efisien dan populer untuk *Independent Component Analysis* yang diciptakan oleh Aapo Hyvärinen di Helsinki University of Technology [10]. *FastICA* dapat secara efektif menghilangkan atau mengurangi efek dari *noise* tanpa memerlukan asumsi-asumsi yang ketat dan tidak realistis [6].

Berdasarkan latar belakang dan rumusan masalah yang telah disampaikan maka tujuan penelitian ini adalah peningkatan akurasi proses analisis sentimen dengan menggunakan metode reduksi dimensi fitur *FastICA* untuk menangani masalah tingginya dimensi dengan data yang non-linier dan tidak terstruktur.

Hasil penelitian ini secara teoritis diharapkan dapat memberikan sumbangan pemikiran dalam bidang text mining atau analisis sentimen dengan integrasi dari algoritma SVM dan reduksi dimensi non-linier *FastICA* yang dapat meningkatkan akurasi model klasifikasi digunakan sebagai metode acuan dalam analisis sentimen dan secara praktis diharapkan dapat membantu pemecahan masalah pihak pengelola pariwisata khususnya Kota Semarang untuk meningkatkan pelayanan, pengelolaan serta fasilitas pendukung pariwisata.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terkait

Penelitian sebelumnya tentang analisis sentimen telah dilakukan oleh Vinodhini dan Chandrasekaran [11] menggunakan analisis komponen utama (PCA) untuk reduksi dimensi dan teknik hibrida untuk klasifikasi sentimen. Mereka melakukan eksperimen dengan 500 sentimen yang terdiri dari 250 positif dan 250 negatif review tentang kamera digital. Untuk representasi fitur, mereka mencoba kombinasi yang berbeda dari *unigram, bi-gram* dan *tri-gram*. SVM dan *Naive Bayes* (NB) dibandingkan terhadap *bagged SVM* and *Bayesian boosting. Bayesian boosting* mengungguli semua metode hibrida (yaitu *unigram + bigram*) representasi fitur dengan menghasilkan akurasi tertinggi 83,3%.

Liu et al. [12] mengembangkan sebuah algoritma dalam dua tahap untuk review kamera. Pada tahap pertama, produk fitur ekstraksi dilakukan menggunakan PMI didasarkan profil dokumen. Untuk prediksi, algoritma *bootstrap* dengan *decision tree* mengungguli MLP, *simple linier regression* (SLR), dan *sequential minimal optimization-support vector regression* (SMOreg). Dalam tahap kedua, analisis komponen utama (PCA), *feature-instance similarity*, dan *Mutual Information* (MI) berdasarkan metode fitur seleksi yang diterapkan untuk mengekstrak fitur untuk menentukan review yang membantu. Eksperimen dilakukan pada 1000 ulasan dikumpulkan dari Amazon.com pada ponsel, dan mencapai MAE 0.599 pada pelatihan tahap pertama, dan MI dan PCA mengungguli untuk 3 dataset.

Deng et al. [13] merancang skema supervised term weighting berdasarkan pentingnya istilah dalam dokumen dan pentingnya istilah untuk mengekspresikan sentimen dengan bantuan tujuh metode seleksi fitur yaitu. *document frequency* (DF), *Information Gain* (IG), *Mutual Information* (MI), *Odds Ratio* (OR), *Chi-Square Statistic* (CHI), *Weighted Log Likelihood Ratio* (WLLR), *Weighted Frequency and Odds* (WFO). Mereka melakukan eksperimen dengan skema pembobotan yang diusulkan menggunakan SVM pada *dataset* Pang et al. [14], *multi-domain dataset*, dan *dataset review* film mencapai akurasi masing-masing 88,5%, 88,7%, dan 88,0%.

2.2. Landasan Teori

2.2.1 Analisis Sentimen

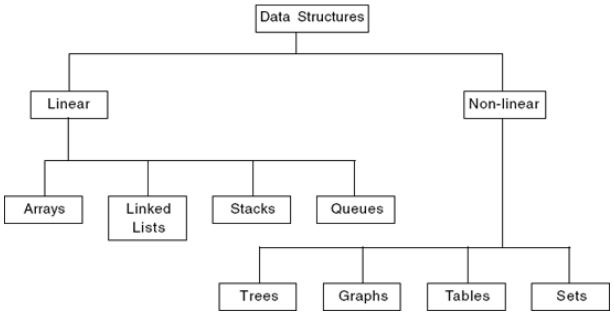
Analisis sentimen merupakan sebuah cabang penelitian di bidang *text mining*. Analisis sentimen atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *text mining*. Analisis sentimen adalah bidang studi yang menganalisis opini seseorang, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa, topik, dan atribut mereka. Ini merupakan ruang masalah besar. Ada juga banyak nama lain dan tugas yang sedikit berbeda, misalnya, analisis sentimen, *opinion mining*, ekstraksi opini, *sentiment mining*, analisis subjektivitas, analisis emosi, *review mining*, dll [1].

2.2.2 Struktur Data

Struktur Data merupakan susunan data dalam memori komputer. Model logis dan matematis dari organisasi data tertentu disebut struktur data. Struktur data meliputi *array, linked list, stack, binary trees*, dan tabel *hash*. Algoritma memanipulasi data dalam struktur ini dengan berbagai cara, seperti mencari *item* data tertentu dan sortir data. Ada dua jenis struktur data yaitu:

a. Struktur data linier
   Setiap struktur data yang mengatur elemen data satu demi satu adalah struktur data linier. Unsur-unsur struktur data linier membentuk urutan atau daftar linier. Dalam struktur data linier, item data disusun dalam urutan linier. Contoh struktur data linier: array, *stack*, antrian, *linked lists*, dll.
b. Struktur data non-linier atau hierarkis
   Struktur data dikatakan non-linier atau hierarkis jika elemen-elemennya tidak membentuk urutan atau daftar linier, sebaliknya strukturnya terlihat hierarkis. Dalam struktur data non-linier, item data tidak berurutan. Contoh struktur data non-linier: *Tree*, Grafik, dan sebagainya.

Struktur data dapat diilustrasikan pada diagram di bawah ini:



Gambar 1. Diagram Struktur Data

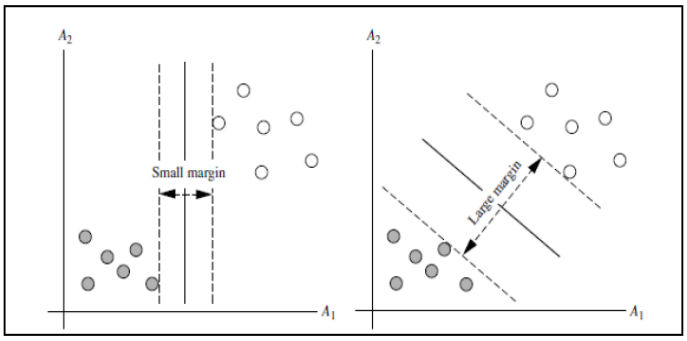
Perbedaan utama antara struktur data linier dan non-linier terletak pada cara mengatur elemen data. Dalam struktur data linier, elemen data disusun secara berurutan dan oleh karena itu mudah diimplementasikan dalam memori komputer. Dalam struktur data non-linier, elemen data bisa dilampirkan ke beberapa elemen data lainnya untuk mewakili hubungan spesifik yang ada diantara mereka. Struktur

data non-linier adalah jika satu elemen dapat dihubungkan ke lebih dari dua elemen yang berdekatan kemudian dikenal sebagai struktur data non-linier.

Dalam domain dunia nyata, data jarang bersifat *linear separable*, kebanyakan bersifat non-linier dan sering memiliki banyak atribut kaya teks atau kata. Misalnya, artikel berita, pengguna di jaringan sosial online seperti *Twitter* dan *Facebook* menerbitkan sejumlah besar status atau review pelanggan untuk sebuah produk atau pelayanan pada berbagai situs web dan lain-lain [19]. Kebanyakan dari fitur dalam data teks dalam domain dunia nyata ini tidak terstruktur dan bersifat non-linier.

2.2.3 *Support Vector Machine (SVM)*

Support Vector Machine membangun *hyperplane* atau himpunan *hyperplane* dalam ruang dimensi tinggi atau tak terbatas, yang dapat digunakan untuk klasifikasi, regresi atau tugas-tugas lainnya. *Hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin *hyperplane* tersebut dan mencari titik maksimalnya. Margin adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing kelas. *Pattern* yang paling dekat ini disebut sebagai *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada *Support Vector Machine* [15].



Gambar 2. Margin Hyperplane SVM

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu $1/\|\vec{w}\|$. Hal ini dapat dirumuskan sebagai *Quadratic Programming (QP) problem*, yaitu mencari titik minimal persamaan, dengan memperhatikan *constraint* persamaan.

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\|\vec{w}\|^2 \quad (1)$$

$$y_i(\vec{x_i}.\vec{w} + b) - 1 \geq 0, \forall i \quad (2)$$

Masalah ini dapat dipecahkan dengan berbagai teknik komputasi, diantaranya *Lagrange Multiplier*

$$L(\vec{w}, b, a) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} a_i \{ y_i((\vec{x_i}.\vec{w} + b) - 1)\} \ (i = 1,2, \ldots, l) \quad (3)$$

a_i adalah *Lagrange multipliers*, yang bernilai nol atau positif ($a_i \geq 0$) . Nilai optimal dari persamaan dapat dihitung meminimalkan L terhadap $\vec{w_l}$ dan *b*, dan dengan memaksimalkan L terhadap a_i . Dengan memperhatikan sifat bahwa pada titik *optimal gradient* L = 0, persamaan dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung a_i saja, sebagaimana persamaan di bawah ini.

$$\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_i \vec{x_i}.\vec{x_j}$$

$$a_i \geq 0 (i = 1,2, \ldots, l) \sum_{i,j=1}^{l} a_i y_i = 0 \quad (4)$$

Dari hasil perhitungan ini diperoleh i yang *ai* kebanyakan bernilai positif. Data yang berkorelasi dengan a_i yang positif inilah yang disebut sebagai *support vector*. Setelah parameter *ai* didapatkan, kemudian masukkan ke persamaan berikut:

$$w = \sum_{i=1}^{N} \alpha_i y_i k(x_i) \quad (5)$$

Hasil yang didapatkan menggunakan persamaan diatas, selanjutnya digunakan untuk mendapatkan nilai *w* dan *b*.

$$y = wx + b \quad (6)$$

Sedemikian sehingga didapatkanlah nilai *w* dan nilai *b* atau nilai *hyperplane* untuk mengklasifikasikan kedua kelas. Dalam SVM terdapat fungsi *kernel* yang dapat mengubah data set yang tidak linier menjadi linier dalam *space* baru. Pemilihan fungsi kernel yang tepat adalah hal yang sangat penting, karena fungsi *kernel* ini akan menentukan *feature space* di mana *classifier* akan dicari.

2.2.4 Independent Component Analysis (ICA)

Untuk memaksimalkan kinerja algoritma klasifikasi SVM maka digunakan *Independent Component Analysis* sebagai metode reduksi dimensi. Analisis komponen Independen (ICA) adalah metode untuk menemukan faktor-faktor yang mendasari atau komponen dari multivarian data statistik (multi-dimensi). Yang membedakan ICA dari metode lain adalah untuk komponen yang keduanya statistik independen, dan *non-gaussian* [16]. Model *Independent Component Analysis* dinyatakan dengan:

$$x = As \quad (7)$$

Huruf *s* adalah vektor sumber yang tidak diketahui dan *A* adalah matriks pencampuran. Sebelum melakukan proses ICA untuk menentukan komponen bebas dari beberapa data, perlu dilakukan penerapan beberapa teknik untuk mendapatkan perhitungan ICA yang efisien, yang biasa disebut dengan *FastICA* (perhitungan ICA dengan hasil yang mendekati kebenaran).

a. Pemusatan Data

Hal yang mendasar yang perlu dilakukan adalah dengan memusatkan data x. Prinsipnya mengurangi data tersebut dengan rerata data dari vektor yang ada dengan Persamaan 8.

$$x_c = x - m \quad (8)$$

sehingga *x* menjadi variabel bererata nol (*zero-mean*). Akibatnya, *s* merupakan variabel yang bererata nol. Pemusatan data ini untuk membuat algoritma ICA menjadi mudah dan cepat. Setelah melakukan taksiran matriks pencampur *A* dengan data memusat, melengkapi taksiran tersebut dengan menambahkan vektor rerata *s* kepada hasil taksiran *s* yang memusat. Vektor rerata *s* adalah hasil kali invers matriks pencampur dan vektor *x* hasil *whitening*, ditunjukkan pada Persamaan 9.

$$s = Wx \quad (9)$$

b. Pemutihan (*Whitening*)

Pemutihan (*whitening*) merupakan praproses yang berfungsi untuk me"mutih"kan variabel yang diamati. Sehingga didapatkan sebuah vektor baru yang variannya sama dengan satu serta komponen nyata dan imajinernya tidak terkorelasi dengan nilai varians yang sama. Secara singkat, matriks kovarians dari $\tilde{X}$ sama dengan matriks identitas (*orthogonal*) dan digambarkan dalam Persamaan 10.

$$E = \{ \tilde{X} \tilde{X}^T \} = I \quad (10)$$

Metode yang sering digunakan adalah dekomposisi nilai *eigen* (*eigen value decomposition – EVD*) dari matriks kovarians $E\{xx^T\}$. Untuk mendapatkan Persamaan 10, dilakukan pemutihan yang dapat dilakukan dengan Persamaan 11.

$$x_w = ED^{-1/2} E^T x \quad (11)$$

D merupakan matriks nilai *eigen* berbentuk diagonal dan E adalah vektor *eigen* dari matriks *x*. Nilai $D^{-1/2} E^T$ disebut sebagai matriks pemutih (*whitening*) dan $ED^{-1/2}$ unuk mengembalikan proses disebut invers matriks pemutih (*dewithening*). Pada tahap ini, terlebih dahulu mengamati nilai dan

vektor *eigen* dari $E\{xx^T\}$, seperti yang dilakukan pada Analisis Komponen Utama (PCA). Persamaan untuk mengamati nilai eigen dirumuskan dengan Persamaan 12.

$$C-Z=|C-\lambda I|$$
$$|C-\lambda I|= 0 \quad (12)$$

C adalah matriks kovarian. *Z* adalah matriks *eigenvalue* dengan λ sebagai *scalar* pembentuknya dan *I* sebagai matriks identitas.

$$|C - \lambda_i I|x = 0 \quad (13)$$

.

2.2.5 *FastICA*

FastICA, sebuah algoritma yang efisien dan populer untuk *Independent Component Analysis* yang diciptakan oleh Aapo Hyvärinen di Helsinki University of Technology [10]. Setelah data mengalami praproses yaitu pemusatan dan pemutihan. Kemudian data diolah dengan suatu metode yang efisien yang disebut *FastICA*. Algoritma *FastICA* dilakukan setiap iterasi. Berikut adalah bentuk algoritma *FastICA* untuk satu unit komponen bebas.

a. Memilih sebuah nilai awal vektor kompleks *w*, dapat secara acak
b. Menghitung nilai *w* yang baru dengan Persamaan 14.

$$w^+ = E\{xg(w_n{}^T x)\} - E\{g'(w_n{}^T x)\}w \quad (14)$$

Fungsi g pada langkah ini adalah fungsi nonlinieritas.

c. Menormalkan nilai *w* yang baru dengan Persamaan 15.

$$W = {w^+}/{\|w^+\|} \quad (15)$$

d. Memeriksa konvergensi, bila tidak konvergen maka kembali ke langkah b.

3. METODE PENELITIAN

Penelitian ini dimulai dari adanya masalah dalam klasifikasi teks pengklasifikasi *Support Vector Machine* (SVM). Pengklasifikasian tersebut memiliki kekurangan terhadap masalah pemilihan parameter yang sesuai, karena dengan tidak sesuainya sebuah pengaturan parameter dapat menyebabkan hasil klasifikasi menjadi rendah. Sumber data yang digunakan dalam penelitian ini yaitu mengambil data status *Twitter* tentang objek wisata di Kota Semarang.

Preprocessing yang dilakukan dengan *case folding, remove punctuation, remove username, remove hashtag, clean number, clean one char, remove url, remove RT, convert number* dan *remove number, remove stop word dan remove emoticon*. Metode pembobotan Fitur yang akan digunakan adalah Term Frequency Invers *Document frequency* (TF-IDF) dan reduksi dimensi fitur menggunakan *FastICA* Sedangkan pengklasifikasi yang digunakan adalah *Support Vector Machine*. Pengujian *10-Fold Cross Validation* akan dilakukan, akurasi algorima akan diukur menggunakan *Confusion Matrix* dan hasil olahan data dam bentuk kurva ROC.

3.1. Pengumpulan Data

Data yang digunakan diambil dari situs jejaring sosial *Twitter*. Satu *tweet* maksimal 140 karakter dapat mewakili sebuah sentimen. Topik dibatasi mengenai objek wisata, proses pencarian menggunakan kata kunci beberapa objek wisata di Kota Semarang. Data terbagi menjadi opini positif dan opini negatif. Opini tersebut ditulis dalam bahasa Indonesia. Sentimen positif yaitu pengguna *Twitter* memberikan komentarnya dengan memberikan penilaian positif terhadap objek wisata tersebut. Sedangkan sentimen negatif merupakan sentimen sebaliknya dari sentimen positif yaitu pengguna memberikan komentarnya dengan memberikan penilaian negatif terhadap objek wisata tersebut.

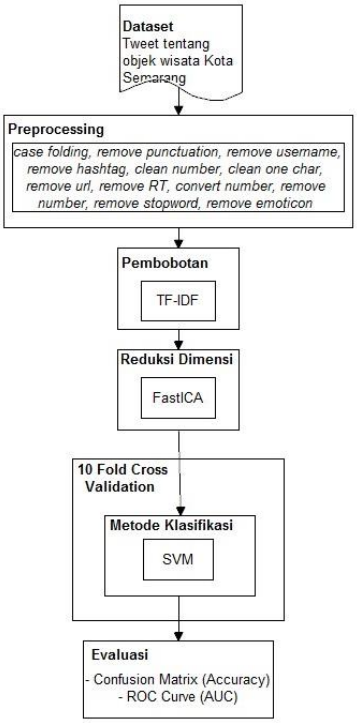
Data diambil pada bulan Januari 2012 – Agustus 2016. Jumlah data *tweet* tentang objek wisata Kota Semarang yang digunakan sebanyak 2000 data, masing-masing 1300 sentimen positif dan 700 sentimen negatif. Data ini termasuk dalam domain dunia nyata (real world problem) yang memiliki banyak *noise*,

tidak terstruktur dan jarang bersifat *linear separable*, kebanyakan bersifat non-linier. Metode yang diusulkan diharapkan dapat mengatasi permasalahan ini. Untuk kelancaran dalam penelitian, alat bantu yang digunakan untuk proses analisis sentimen yaitu perangkat lunak Matlab 2016 dan Weka 3.8.

3.2. Proses Analisis Sentimen

Pada penelitian ini metode yang diusulkan dimulai dengan melakukan *Preprocessing* yang dilakukan dengan teknik *case folding, remove punctuation, remove username, remove hashtag, clean number, clean one char, remove url, remove RT, convert number* dan *remove number, remove stop word dan remove emoticon*. Metode pembobotan Fitur yang akan digunakan adalah *Term Frequency Invers Document frequency* (TF-IDF) dan reduksi dimensi fitur menggunakan *FastICA* untuk mengatasi masalah tingginya dimensi fitur. Sedangkan pengklasifikasi yang digunakan adalah *Support Vector Machine*. Pengujian *10-Fold Cross Validation* akan dilakukan, akurasi algorima akan diukur menggunakan *Confusion Matrix* dan hasil olahan data dam bentuk kurva ROC.

Berikut ini adalah metode yang digunakan untuk proses analisis sentimen dalam penelitian ini.



Gambar 3. Proses Analisis Sentimen

3.3. Pembobotan *Term*

Proses *text mining* dimulai dengan mengubah teks menjadi data vektor. Vektor dalam penelitian ini memiliki dua komponen yaitu dimensi (*word id*) dan bobot. Pembobotan *term* merupakan *term documents matrix* yang representasi kumpulan dokumen yang digunakan untuk melakukan proses klasifikasi dokumen teks. Pada penelitian ini akan digunakan metode TF-IDF sebagai proses pembobotan, yaitu akan dilakukan pembobotan pada tiap *term* berdasarkan tingkat kepentingan di dalam sekumpulan dokumen masukan. Tujuan dari model ruang vektor digunakan untuk memberikan setiap kata dalam dokumen sebuah ID (dimensi) dan sebuah bobot berdasarkan seberapa penting keberadaannya dalam dokumen (posisi dokumen dalam dimensi itu). Adapun perhitungan bobot yang digunakan adalah:

$$TF - IDF = df_i . \log D / df_i \quad (16)$$

- df_i adalah banyaknya dokumen yang mengandung fitur *i* (kata) yang dicari
- D adalah jumlah dokumen

3.4. Reduksi Dimensi Fitur

Dalam penelitian ini digunakan metode reduksi fitur non-linier, yaitu *FastICA* yang dapat menangani masalah pada data domain dunia nyata yang bersifat tidak terstruktur dan non-linier sehingga bisa meningkatkan akurasi *Support Vector Machine*. Sebelum melakukan proses ICA untuk menentukan komponen bebas dari beberapa data, perlu dilakukan *Preprocessing* seperti *centering* dan *whitening* untuk meningkatkan ketepatan algoritma ICA dengan cara mengurangi jumlah *Independent Component* yang harus diestimasi dan menurunkan tingkat kompleksitas data [16], yang kemudian dilanjutkan dengan metode reduksi dimensi non-linier *FastICA* (perhitungan ICA dengan hasil yang mendekati kebenaran), algoritma *FastICA* dilakukan pada setiap iterasi.

3.5. Metode Klasifikasi

Proses klasifikasi di sini adalah untuk menentukan sebuah kalimat sebagai anggota kelas positif atau kelas negatif berdasarkan nilai perhitungan. Teknik klasifikasi yang akan digunakan untuk penelitian ini adalah *Support Vector Machine* (SVM). SVM memiliki kelebihan yaitu mampu mengidentifikasi *hyperplane* terpisah yang memaksimalkan margin antara dua kelas yang berbeda [17]. Seperti yang diketahui data yang digunakan dalam penelitian ini bersifat *non linear separable*, oleh karena itu digunakan fungsi *kernel polynomial* pangkat 1 yang didefiniskan sebagai $(x_i,)=(x_i+x_i^T+1)^1$. Reduksi dimensi fitur sekaligus pengaturan parameter pada SVM secara signifikan dapat mempengaruhi hasil akurasi klasifikasi.

Penelitian ini menghasilkan klasifikasi teks dalam bentuk positif atau negatif. Pengukuran berdasarkan akurasi *Support Vector Machine* sebelum dan sesudah pembobotan dan reduksi dimensi fitur.

3.6. Validasi & Evaluasi

Validasi yang digunakan dalam penelitian ini adalah *k-fold cross validation*. Dalam *k-fold cross validation*, data awal dipartisi secara acak menjadi sejumlah k subset (*fold*), yaitu D_1, D_2, ..., D_k, yang masing-masingnya berukuran sama. Cara kerja *K-fold cross validation* adalah sebagai berikut:

a. Total *instance* dibagi menjadi N bagian.
b. *Fold* ke-1 adalah ketika bagian ke-1 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut. Perhitungan akurasi tersebut dengan menggunakan persamaan sebagai berikut:

$$Akurasi = \frac{\sum data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} \times 100\% \quad (17)$$

c. *Fold* ke-2 adalah ketika bagian ke-2 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.
d. Demikian seterusnya hingga mencapai fold ke-K. Hitung rata-rata akurasi dari K buah akurasi di atas. Rata-rata akurasi ini menjadi akurasi final.

Validasi menggunakan 10 *Fold Cross Validation* dimana data dibagi secara acak menjadi 10 bagian data dengan jumlah yang sama. Sehingga dilakukan proses validasi sebanyak 10 kali secara berulang.

4. HASIL PENELITIAN DAN PEMBAHASAN

4.1. Klasifikasi tanpa Reduksi Dimensi

Pada tahapan eksperimen ini dilakukan pengklasifikasian sentimen menggunakan algoritma SVM. Pengukuran dengan *Confusion Matrix* di sini akan menampilkan hasil akurasi model SVM sebelum ditambahkan metode reduksi dimensi *FastICA* yang bisa dilihat pada Tabel 1.

Tabel 1. *Confusion Matrix* Algoritma SVM Sebelum Penambahan Metode Reduksi Dimensi

Akurasi: 91.95 %		***Actual Class***	
		true positif	*true negatif*
Prediction class	*Prediction positif*	1245	55
	Prediction negatif	106	594

Perhitungan akurasi dari tabel *Confusion Matrix* tersebut adalah sebagai berikut:

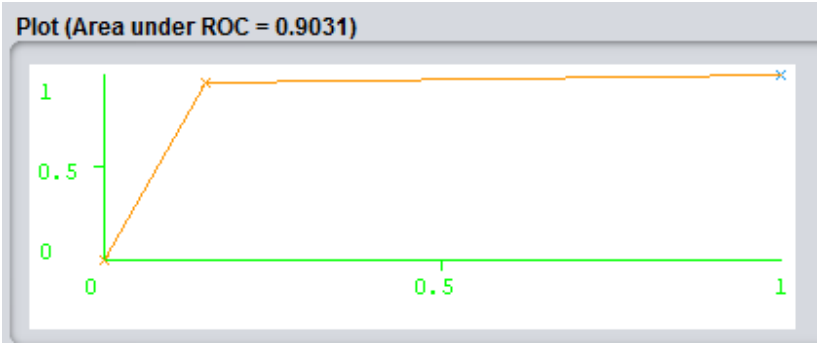
$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$akurasi = \frac{1245 + 594}{1245 + 594 + 106 + 55} \times 100\%$$

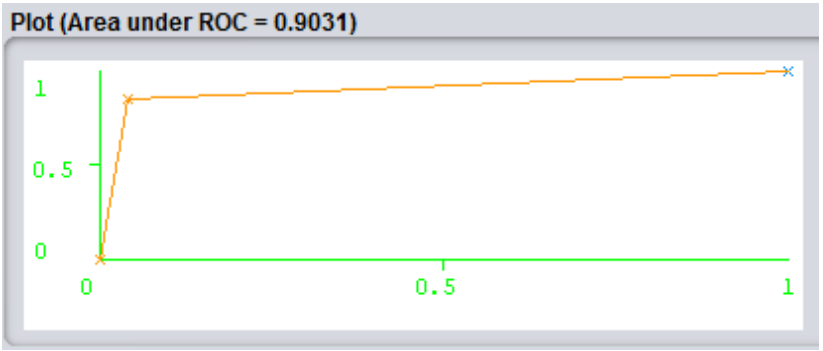
$$akurasi = 91.95\ \%$$

Dari sebanyak 2.000 dataset *Twitter* tentang objek wisata Kota Semarang yaitu 1.300 sentimen positif dan 700 sentimen negatif, sebanyak 594 data diprediksi sesuai yaitu negatif, dan sebanyak 106 data diprediksi negatif tetapi ternyata positif, 1245 data diprediksi sesuai yaitu positif dan 55 data diprediksi positif tetapi ternyata negatif. Hasil yang diperoleh dengan menggunakan algoritma SVM dengan nilai akurasi 91.95 % .

Hasil perhitungan divisualisasikan dengan kurva ROC bisa dilihat pada gambar di bawah, pada bagian atas dari grafik juga ditampilkan nilai dari AUC (*Area Under Curve*).



Gambar 4. Kurva ROC *Support Vector Machine* Kelas Positif



Gambar 5. Kurva ROC *Support Vector Machine* Kelas Negatif

Dari Gambar 5 dan Gambar 6 terdapat grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.9031 dimana diagnosa hasil klasifikasi sangat baik (*excellent classification*). Untuk mencapai nilai keakurasian AUC mendekati 1 (sangat sempurna) dibutuhkan metode untuk meningkatkan diagnosa hasil klasifikasi yang terbentuk. Dalam hal ini peneliti menggunakan *FastICA* sebagai algoritma reduksi dimensi fitur untuk meningkatkan akurasi klasifikasi.

4.2. Klasifikasi dengan Reduksi Dimensi

4.2.1 Klasifikasi dengan Metode Reduksi Dimensi *FastICA*

Pada tahapan eksperimen ini dilakukan pengklasifikasian sentimen menggunakan algoritma SVM dengan metode reduksi dimensi *FastICA* untuk mendapatkan hasil klasifikasi yang optimal dengan cara pemilihan parameter yang berbeda pada algoritma *FastICA*. Parameter yang ditentukan adalah jumlah dimensi hasil reduksi menggunakan algoritma *FastICA*. Berikut ini adalah tabel hasil perbandingannya:

Tabel 2. Perbandingan
Hasil Klasifikasi Berdasakan Jumlah Dimensi *FastICA*

Jumlah dimensi	Akurasi	AUC
100	90.50 %	0.8791
200	90.55 %	0.8828
300	91.15 %	0.8943
400	92.40 %	0.9096
410	92.30 %	0.9078
420	92.90 %	0.9157
430	92.10 %	0.9073
440	92.35 %	0.9108
450	92.10 %	0.9073
500	77.15 %	0.7286
600	74.50 %	0.6998
700	70.30 %	0.6400

Dapat dilihat pada Tabel 2 didapatkan akurasi tertinggi 92.90 % pada jumlah dimensi 420 sehingga dapat disimpulkan itulah parameter terbaik yang dapat digunakan pada penelitian ini. Pengukuran dengan *Confusion Matrix* di sini akan menampilkan hasil akurasi klasifikasi dengan algoritma SVM sesudah ditambahkan metode reduksi dimensi *FastICA* dengan jumlah dimensi 420 yang bisa dilihat pada Tabel 3.

Tabel 3. *Confusion Matrix*
Algoritma SVM Sesudah Penambahan Metode *FastICA*

Akurasi: 92.90 %		***Actual Class***	
		true positif	*true negatif*
Prediction class	*Prediction positif*	1248	52
	Prediction negatif	90	610

Perhitungan akurasi dari tabel *Confusion Matrix* tersebut adalah sebagai berikut:

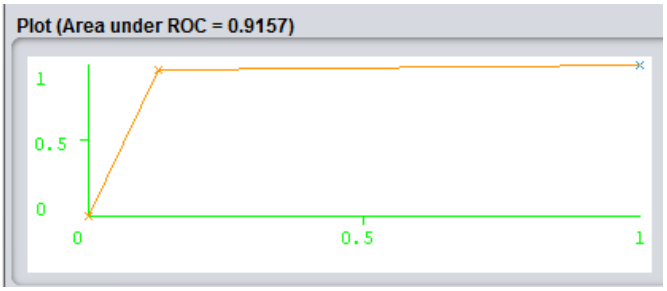
$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$akurasi = \frac{1248 + 610}{1248 + 610 + 90 + 52} \times 100\%$$

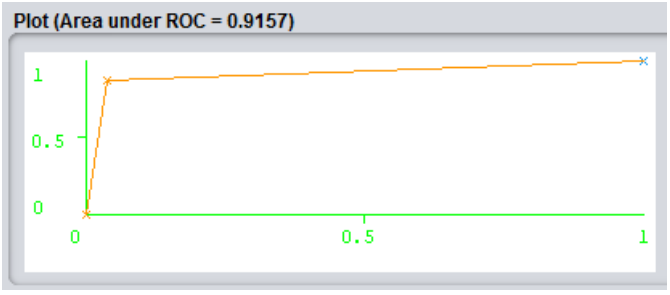
$$akurasi = 92.90\ \%$$

Dari sebanyak 2.000 dataset *Twitter* objek wisata Kota Semarang yaitu 1.300 sentimen positif dan 700 sentimen negatif, sebanyak 610 data diprediksi sesuai yaitu negatif, dan sebanyak 90 data diprediksi negatif tetapi ternyata positif, 1248 data diprediksi sesuai yaitu positif dan 52 data diprediksi positif tetapi ternyata negatif. Hasil pengujian *Confusion Matrix* di atas diketahui bahwa menggunakan algoritma SVM mempunyai akurasi hanya 91.95 % sedangkan algoritma SVM dengan metode reduksi dimensi *FastICA* memiliki tingkat akurasi yang lebih tinggi yaitu 92.90%. Akurasi naik 0.95 % dari yang sebelumnya.

Hasil perhitungan divisualisasikan dengan kurva ROC yang bisa dilihat pada gambar di bawah, pada bagian atas dari grafik juga ditampilkan nilai dari AUC (*Area Under Curve*).



Gambar 6. Kurva ROC *Support Vector Machine* dengan *FastICA* Kelas Positif



Gambar 7. Kurva ROC *Support Vector Machine* dengan *FastICA* Kelas Negatif

Dari Gambar 6 dan Gambar 7 terdapat kurva ROC SVM dengan penambahan reduksi dimensi *FastICA* dengan nilai AUC (Area Under Curve) sebesar 0.9157 yang menunjukkan diagnosa hasil klasifikasi sangat baik (*excellent classification*).

4.2.2 Klasifikasi dengan Metode Reduksi Dimensi PCA

Pada penelitian terkait sebelumnya [11][12] digunakan *Principal Component Analysis* (PCA) sebagai metode reduksi dimensi. Sehingga pada tahap ini dilakukan juga klasifikasi dataset *Twitter* objek wisata Kota Semarang dengan menggunakan metode PCA sehingga hasilnya dapat dibandingkan dengan metode yang diusulkan. Parameter yang ditentukan adalah jumlah dimensi hasil reduksi menggunakan algoritma PCA. Berikut ini adalah tabel hasil perbandingannya.

Tabel 4. Perbandingan

Hasil Klasifikasi Berdasakan Jumlah Dimensi PCA

Jumlah dimensi	Akurasi	AUC
100	91.04 %	0.8979
200	91.70 %	0.8973
250	92.10 %	0.9026
260	92.30 %	0.9045
270	92.00 %	0.9012
300	91.55 %	0.8991
400	91.40 %	0.8979
500	91.10 %	0.8946
600	90.00 %	0.8832
700	88.90 %	0.8698

Dapat dilihat pada Tabel 4 didapatkan akurasi tertinggi 92.30 % pada jumlah dimensi 260 sehingga dapat disimpulkan itulah parameter terbaik yang dapat digunakan. Pengukuran dengan *Confusion Matrix* selanjutnya akan menampilkan perbandingan dari hasil akurasi klasifikasi dengan algoritma SVM sesudah ditambahkan metode reduksi dimensi PCA dengan jumlah dimensi 260 yang bisa dilihat pada Tabel 5.

Tabel 5. *Confusion Matrix* Algoritma SVM Sesudah Penambahan Metode PCA

Akurasi: 92.30 %		***Actual Class***	
		true positif	*true negatif*
Prediction class	*Prediction positif*	1256	44
	Prediction negatif	110	590

Perhitungan akurasi dari tabel *Confusion Matrix* tersebut adalah sebagai berikut:

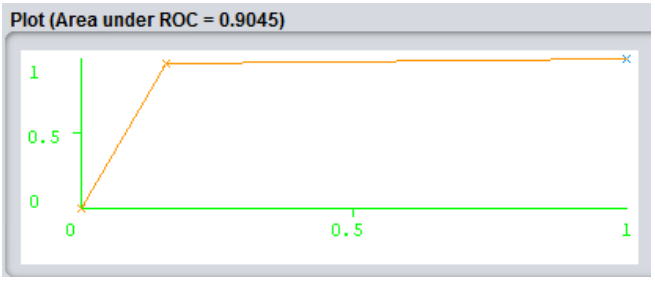
$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$akurasi = \frac{1256 + 590}{1256 + 590 + 110 + 44} \times 100\%$$

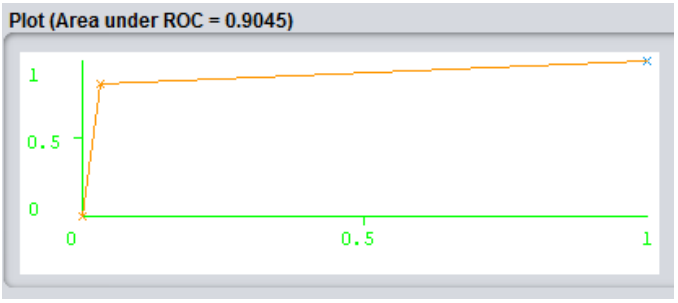
$$akurasi = 92.30\ \%$$

Hasil pengujian *Confusion Matrix* di atas diketahui bahwa menggunakan algoritma SVM mempunyai akurasi hanya 91.95 % sedangkan algoritma SVM dengan metode reduksi dimensi PCA memiliki tingkat akurasi yang lebih tinggi yaitu 92.30%.

Hasil perhitungan divisualisasikan dengan kurva ROC yang bisa dilihat pada gambar di bawah, pada bagian atas dari grafik juga ditampilkan nilai dari AUC (*Area Under Curve*).



Gambar 8. Kurva ROC *Support Vector Machine* dengan PCA Kelas Positif



Gambar 9. Kurva ROC *Support Vector Machine* dengan PCA Kelas Negatif

Dari Gambar 8 dan Gambar 9 terdapat kurva ROC SVM dengan penambahan reduksi dimensi PCA dengan nilai AUC (Area Under Curve) sebesar 0.9045 yang menunjukkan diagnosa hasil klasifikasi baik (*good classification*).

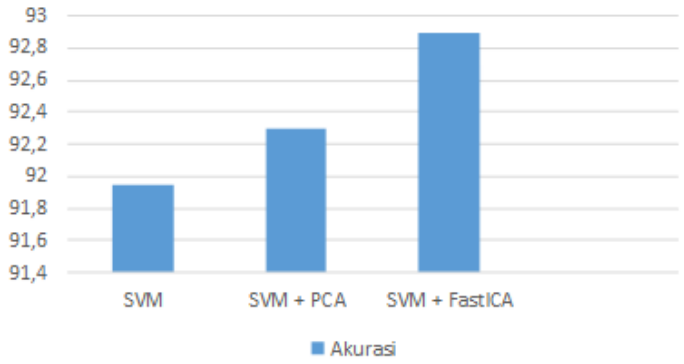
4.2.3 Perbandingan Hasil Klasifikasi

Dari beberapa tahapan eksperimen yang telah dilakukan di atas dapat dibuat suatu perbandingan dari hasil masing-masing model terhadap dataset *Twitter* tentang objek wisata Kota Semarang. Dalam penelitian ini, menunjukkan seberapa baik model yang terbentuk. Tanpa menggunakan metode reduksi fitur, algoritma SVM sendiri sudah menghasilkan akurasi sebesar 91.95 % dan nilai AUC 0.9031. Akurasi tersebut masih kurang akurat, sehingga perlu ditingkatkan lagi menggunakan metode reduksi dimensi fitur *FastICA*. Setelah menggunakan metode reduksi dimensi *FastICA*, akurasi algoritma SVM meningkat menjadi 92.90% dan nilai AUC 0.9157 namun dengan menggunakan metode reduksi dimensi PCA akurasi menurun menjadi 92.30 % dan nilai AUC 0.9045 seperti yang bisa dilihat pada tabel berikut:

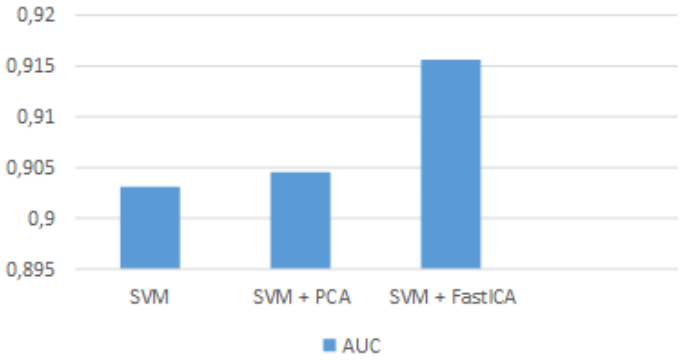
Tabel 6. Perbandingan Model Algoritma SVM
Sebelum dan Sesudah Menggunakan Metode Reduksi Dimensi

	Algoritma SVM	**Algoritma SVM + PCA**	**Algoritma SVM + *FastICA***
Sukses prediksi positif	1245	1256	1248
Sukses prediksi negatif	549	590	610
Akurasi Model	91.95 %	92.30 %	92.90%
AUC	0.9031	0.9045	0.9157

Berdasarkan hasil evaluasi di atas diketahui bahwa algoritma SVM yang menggunakan reduksi dimensi fitur *FastICA*, mampu meningkatkan tingkat akurasi analisis sentimen objek wisata Kota Semarang daripada kedua metode lainnya. Gambar 10 memperlihatkan perbandingan tingkat akurasi yang meningkat dalam bentuk sebuah grafik. Sedangkan Gambar 11 memperlihatkan perbandingan nilai AUC.



Gambar 10. Grafik Perbandingan Akurasi



Gambar 11. Grafik Perbandingan Nilai AUC

Nilai AUC menunjukkan hasil yang lebih baik jika nilainya mendekati 1. Nilai AUC pada dataset juga mengalami peningkatan daripada menggunakan metode SVM tanpa reduksi dimensi dan SVM menggunakan reduksi dimensi PCA. Oleh karena itu, dapat disimpulkan bahwa metode reduksi dimensi *FastICA*, terbukti dapat meningkatkan akurasi algoritma SVM pada data *Twitter* objek wisata Kota Semarang sehingga dapat diklasifikasi dengan baik ke dalam kelas positif dan negatif.

5. PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dari tahap analisa permasalahan dan *literature review* sampai dengan tahap evaluasi hasil, telah dapat diketahui hasil evaluasi dari metode yang diusulkan. Pengujian model dengan menggunakan data non-linier *tweet* tentang objek wisata di Kota Semarang untuk sentimen positif maupun negatif dengan keseluruhan data 2000 *tweet* atau dokumen tidak terstruktur, bersifat *non linear separable* dan memiliki ribuan fitur yang termasuk dalam dimensi tinggi.

Pada algoritma klasifikasi *Support Vector Machine* digunakan fungsi *kernel polynomial* pangkat 1 untuk mengatasi masalah data non-linier, matrik kernel ini digunakan untuk membuat data yang bersifat non-linear menjadi linear sehingga bisa mendapatkan fungsi pemisah linier di dalam *feature space*. Model yang dihasilkan diuji mendapatkan nilai akurasi dan nilai AUC dari setiap algoritma sehingga didapatkan pengujian dengan menggunakan *Support Vector Machine* didapatkan nilai akurasi 91.95 % dan nilai AUC 0.9031. Sedangkan pengujian dengan menggunakan *Support Vector Machine* dengan penambahan metode reduksi dimensi non-linier *FastICA* didapatkan nilai akurasi 92.90% dan nilai AUC 0.9157 yang berarti akurasi naik 0.95 % lebih baik dari pada *Support Vector Machine* sendiri.

Dari hasil evaluasi di atas, dapat disimpulkan bahwa penggunaan metode reduksi dimensi *FastICA* untuk mengatasi masalah dimensi tinggi pada dataset dengan data non-linier dan dapat menghasilkan analisis sentimen yang lebih akurat.

5.2. Saran

a. Disarankan untuk menggunakan metode reduksi dimensi fitur yang lain, seperti *Fisher's linier Discrimination Ratio, Latent Semantic Indexing* dan lain-lain dikombinasikan dengan metode pengklasifikasi lain yang mungkin di luar *Supervised Learning* sebagai pengembangan dari penelitian ini.
b. Metode SVM pada penelitian ini adalah pengklasifikasian biner yang hanya menghasilkan dua kelas. Selanjutnya dapat diteliti implementasi dan unjuk kerja metode SVM untuk pengklasifikasian teks *multiclass*.

UCAPAN TERIMAKASIH

Penulis mengucapkan banyak terima kasih kepada :
a. Bapak Heru Agus Santoso, Ph. D dan Bapak Catur Supriyanto, M. Cs selaku pembimbing yang telah meluangkan waktu serta memberi saran dan dukungan selama penyusunan penelitian ini.
b. Seluruh staf pengajar Universitas Dian Nuswantoro Semarang yang telah membimbing dan membagi ilmu selama masa perkuliahan, sehingga dapat dipergunakan sebagai dasar pertimbangan dalam penyusunan penelitian ini.
c. Orang tua saya yang selalu memberi motivasi, nasihat dan mendoakan saya.

PERNYATAAN ORISINALITAS

" Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya"
[**Mochamad Amry Assiva**]

DAFTAR PUSTAKA

[1] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[2] D. J. Harper and M. F. Porter, "The Selection of Good Search Terms ?," vol. 16, no. 6, pp. 271–285.
[3] C. Lee and G. G. Lee, "*Information Gain* and divergence-based feature selection for machine learning-based text categorization q," vol. 42, pp. 155–165, 2006.
[4] P. Magdalinos, C. Doulkeridis, and M. Vazirgiannis, "Enhancing Clustering Quality through Landmark-Based Dimensionality Reduction," *ACM Trans. Knowl. Discov. Data*, vol. 5, no. 2, pp. 1–44, 2011.
[5] W. Fan and W. Xi, "Effective and efficient dimensionality reduction for large-scale and streaming data *Preprocessing*," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 3, pp. 320–333, 2006.
[6] X. Tian, L. Cai, and S. Chen, "Neurocomputing Noise-resistant joint diagonalization independent component analysis based process fault detection," *Neurocomputing*, vol. 149, pp. 652–666, 2015.
[7] E. Alpaydın, *Introduction to machine learning*, vol. 1107. 2014.
[8] I. a Practical, M. Data, A. D. P. Berrar, and W. Dubitzky, "Chapter 5 Singular value decomposition and principal component analysis," *A Pract. approach to microarray data Anal.*, vol. 44, no. 2, pp. 1–18, 2003.
[9] M. Rezghi and A. Obulkasim, "Noise-free principal component analysis: An efficient dimension reduction technique for high dimensional molecular data," *Expert Syst. Appl.*, vol. 41, no. 17, pp. 7797–7804, 2014.

[10] A. Hyvarinen, "Fast and robust fixed-point algorithm for independent component analysis," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 10, pp. 626–634, 1999.
[11] G. Vinodhini and R. M. Chandrasekaran, "Opinion mining using principal component analysis based ensemble model for e-commerce application," *CSI Trans. ICT*, vol. 2, no. 3, pp. 169–179, 2014.
[12] Y. Liu, J. Jin, P. Ji, J. A. Harding, and R. Y. K. Fung, "Identifying helpful online reviews: A product designer's perspective," *CAD Comput. Aided Des.*, vol. 45, no. 2, pp. 180–194, 2013.
[13] Z. H. Deng, K. H. Luo, and H. L. Yu, "A study of supervised term weighting scheme for sentiment analysis," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3506–3513, 2014.
[14] J. Prager, "Open-Domain Question–Answering," *Found. Trends® Inf. Retr.*, vol. 1, no. 2, pp. 91–231, 2006.
[15] R. G. Brereton and G. R. Lloyd, "*Support Vector Machine*s for classification and regression.," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.
[16] A. rinen, "Independent component analysis, algorithms and applications," *Neural Networks*, vol. 136, no. 1, pp. 411–430, 2000.
[17] T. Pang-Ning, M. Steinbach, and V. Kumar, "Introduction to data mining," *Libr. Congr.*, p. 796, 2006.
[18] D. Prasetya and M. Rani, "Pengembangan Potensi Pariwisata Kabupaten Sumenep , Madura , Jawa Timur (Studi Kasus: Pantai Lombang)," vol. 3, no. 3, pp. 412–421, 2014.
[19] F. S. Gharehchopogh, "Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing," pp. 1–4, 2011.