Klasifikasi Konten Berita Digital Bahasa Indonesia Menggunakan Support Vector Machines (SVM) Berbasis Particle Swarm Optimization (PSO)

Acmad Nurhadi [1)]
[1)] Akademi Manajemen Informatika dan Komputer "BSI Pontianak"
Email : achmad.ahh@bsi.ac.id

Abstrak - Banyak instansi yang bergerak dalam penyaluran informasi atau berita sudah mulai menggunakan sistem berbasis *web* untuk menyampaikan berita secara *up to date.* Namun, dalam membagi berita ke dalam kategori-kategori tersebut untuk saat ini masih dilakukkan secara manual, sehingga sangat merepotkan dan juga dapat memakan waktu yang lama. Dari beberapan teknik tersebut yang paling sering digunakan untuk klasifikasi konten berita adalah *Support Vector Machine* (SVM). SVM memiliki kelebihan yaitu mampu mengidentifikasi *hyperplane* terpisah yang memaksimalkan margin antara dua kelas atau lebih yang berbeda. Pemilihan fitur sekaligus penyetingan parameter di SVM secara signifikan mempengaruhi hasil akurasi klasifikasi. Oleh karena itu, dalam penelitian ini digunakan penggabungan metode pemilihan fitur, yaitu *Particle Swarm Optimization* agar bisa meningkatkan akurasi pengklasifikasi *Support Vector Machine.* Penelitian ini menghasilkan klasifikasi teks dalam bentuk kategori gosip, kuliner, dan travel dari konten berita digital. Pengukuran berdasarkan akurasi *Support Vector Machine* sebelum dan sesudah penambahan metode pemilihan fitur. Evaluasi dilakukan menggunakan 10 *fold cross validation.* Sedangkan pengukuran akurasi diukur dengan *confusion matrix.* Hasil penelitian menunjukan peningkatan akurasi *Support Vector Machine* dari 65.81% menjadi 95.42%

Kata kunci : Particle Swarm Optimization, Support Vector Machine, Klasifikasi Konten Berita, Text Mining

1. PENDAHULUAN

Berita merupakan informasi baru atau informasi mengenai sesuatu yang sedang terjadi, disajikan lewat bentuk cetak, siaran, internet, atau dari mulut ke mulut kepada orang ketiga atau orang banyak. Di era perkembangan teknologi ini, berita dapat dilihat menggunakan internet seperti kompas.com yang merupakan salah satu website berita yang sering dikunjungi. Banyak informasi yang dapat kita terima dalam website tersebut. Terkadang, kita langsung saja menerima tanpa adanya penyeleksian informasi. Atas dasar itu banyak dari media informasi dimasa sekarang yang melakukan pengklasifikasian dengan kategorisasi terlebih dahulu sebelum disebarkan pada masyarakat luas. Pengklasifikasian tersebut berguna untuk memudahkan masyarakat untuk mencari informasi yang mereka inginkan.

Untuk mempermudah dalam pengklasifikasian, dengan menggunakan metode text mining sebagai salah satu alternatif untuk menyelesaikannya. Text mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks. Proses penganalisisan teks ini berguna untuk mencari informasi yang bermanfaat untuk tujuan tertentu.

Ada beberapa penelitian yang sudah dilakukan dalam melakukkan klasifikasi konten berita diantaranya, sebuah pendekatan baru *text mining* berdasarkan HMM-SVM untuk klasifikasi berita pada *web* (Krishnalal, Rengarajan, dan Srinivasagan, 2010). *Automated* kategorisasi teks arab menggunakan SVM dan NB (Alsaleem, 2011). Efektivitas kategorisasi otomatis dokumen thailand berdasarkan *machine learning* (Janpla, 2014). Teks kategorisasi Arab menggunakan regresi logistik (Tahrawi, 2015). Dan penelitian tentang mengoptimalkan parameter *Support Vector Machine* menggunakan algoritma genetika untuk klasifikasi kecenderungan sengketa di tahap awal proyek kemitraan publik-swasta (Chou et al., 2014).

Dari beberapa penelitian tersebut, teknik yang banyak digunakan untuk klasifikasi data adalah *Support Vector Machines* (SVM). SVM dipertahankan untuk melakukkan yang terbaik, dan unigram dengan keberadaan informasi perubahan keluar untuk menjadi fungsi yang paling bermanfaat (Basari et al., 2013). SVM memiliki kelebihan yaitu mampu mengidentifikasikan *hyperplane* terpisah yang memaksimalkan margin antara dua kelas yang berbeda (Chou et al., 2014). Namun, SVM memiliki kekurangan terhadap masalah pemilihan parameter atau fitur yang sesuai (Basari et al., 2013).

Ada beberapa teknik yang bisa digunakan untuk lebih menyempurnakan kekurangan SVM tersebut, yaitu salah satu nya menggunakan *Particle Swarm Optimization* (PSO). *Particle Swarm Optimization* (PSO) banyak digunakan untuk memecahkan masalah optimasi serta sebagai masalah seleksi fitur (Liu et al., 2011). Selain itu *Particle Swarm Optimization* (PSO) adalah suatu teknik optimasi yang sangat sederhana untuk

menerapkan dan memodifikasi beberapa parameter (Basari et al., 2013).

Pada penelitian ini algoritma *Support Vector Machines* dan algoritma *Particle Swarm Optimization* sebagai metode pemilihan fitur akan diterapkan oleh penulis untuk mengklasifikasikan teks pada isi konten berita digital bahasa Indonesia untuk mengelompokkan isi konten berita tersebut sesuai dengan kategorinya masing-masing.

2.1. Tinjauan Studi

Ada beberapa penelitian yang menggunakan *Support Vector Machines* sebagai pengklasifikasi dalam klasifikasi teks isi konten berita digital, diantaranya:

2.1.1. Model Penelitian Bambang

Penelitian yang dilakukkan oleh Bambang Kurniawan, Syahril Effendi dan Opim Salim Sitompul mengenai klasifikasi konten berita dengan metode *text mining*. Konten berita yang dijadikan *dataset* diambil dari beberapa berita *online* yang terdiri dari 4 kategori yaitu berita ekonomi, berita olahraga, berita politik dan berita *entertainment.* Langkah-langkah yang dilakukkan terdiri dari pengumpulan data, *text mining* atau menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar, setelah itu tahap *preprocessing* yang terdiri dari *toLowerCase, Tokenizing, Feature Selection, stopword removal, stemming* dan terakhir menggunakan algoritma *Confix-stripping Stemmer* (Kurniawan, Effendi dan Sitompul, 2012).

2.1.2. Model Penelitian Krishnalal

Penelitian yang dilakukkan oleh Krishnalal G, S Babu Rengarajan dan KG Srinivasagan mengenai sebuah pendekatan baru *text mining* berdasarkan HMM-SVM untuk klasifikasi berita pada *web.* Konten berita yang dijadikan *dataset* terdiri dari 1.200 konten berita yang diambil dari situs www.hinduonnet.com, http://timesofindia.com, www.expressbuzz.com, www.thehindubusinessline.com. Langkah-langkah yang dilakukkan terdiri dari proses *preprocessor* yaitu dilakukkan dengan input isi konten berita dalam bentuk file teks dengan menggunakan JfileChooser API yang kemudian mengalami Parsing atau membagi teks ke satu set bagian diskrit atau token, lalu setelah itu menghilangkan tanda baca, angka, karakter bukan abjad, *stopwords* dan proses *stemming* yaitu penghilangan awalan, akhiran dan sebagainya. Langkah selanjutnya yaitu proses klasifikasi menggunakan pengklasifikasi Hidden Markov Model dan *Support Vector Machines*.

Dari penelitian tersebut, terbukti *Support Vector Machines* memiliki tingkat akurasi lebih tinggi dibandingkan dengan pengklasifikasi lainnya (Krishnalal, Rengarajan, dan Srinivasagan, 2010).

2.1.3. Model Penelitian Alsaleem

Penelitian yang dilakukan oleh Saleh Alsaleem mengenai kategorisasi otomatis teks bahasa arab menggunakan SVM dan NB. *Dataset* diambil dari Saudi *Newspapers* yang terdiri dari 51521 dokumen Arab yang dibagi menjadi 7 kategori terdiri dari Budaya, Ekonomi, General, Teknologi Informasi, Politik, Sosial dan Olahraga. Langkah-langkah yang dilakukkan terdiri dari Setiap artikel dalam kumpulan data Arab diproses untuk menghapus angka dan tanda baca, setelah itu melakukkan normalisasi beberapa huruf Arab seperti normalisasi (hamza (ا) atau (إ)) dalam segala bentuknya ke (alef (ا)). Semua teks Arab non disaring. kata-kata fungsi Arab telah dihapus. Kata-kata fungsi Arab (menghentikan kata-kata) adalah kata-kata yang tidak berguna dalam sistem IR misalnya Arab prefiks, kata ganti, dan preposisi. Langkah selanjutnya yaitu proses klasifikasi menggunakan *Support Vector Machines* dan Naïve Bayes.

Dari penelitian tersebut, *Support Vector Machines* mengungguli NB pada enam set data berkaitan dengan hasil F1. Hasil presisi mendapatkan bahwa SVM mengungguli NB pada empat set (Alsaleem, 2011).

2.1.4. Model Penelitian Janpla

Penelitian yang dilakukkan oleh Satien Janpla mengenai efektifitas kategorisasi otomatis dokumen thai berdasarkan machine learning. *Dataset* yang diambil dari dokumen elektronik Thailand dan layanan berita *online* melalui internet dalam 10 kelompok terdiri dari 12.000 dokumen yang meliputi pendidikan, hiburan, sosial, politik, teknologi, olahraga, asing, pertanian, ekonomi dan Bangkok. Langkah-langkah yang dilakukkan terdiri dari melakukkan *feature extraction* mengubah dokumen ke format yang bisa digunakan oleh komputer, lalu setelah itu tahap *stopwords* atau memotong kata-kata yang tidak perlu, lalu tahap *word segmentation, indexing, model* dan *performance model.* Langkah selanjutnya proses pengklasifikasian menggunakan *Support Vector Machines*, *Decision Trees* dan Naïve Bayes.

Dari penelitia, terbukti *Support Vector Machines* memiliki tingkat akurasi lebih tinggi dibandingkan dengan pengklasifikasi lainnya dalam hal ini Decision Trees dan Naïve Bayes (Janpla, 2014).

2.1.5. Model Penelitian Al-Tahrawi

Penelitian yang dilakukan Mayy M. Al-Tahrawi membahas tentang teks kategorisasi Arab menggunakan regresi logistik. *Dataset* yang diambil dari 1500 artikel berita arab yang dibagi menjadi 5 kategori yaitu seni, ekonomi, politik, sains dan olahraga. Langkah-langkah yang dilakukan dalam penelitian tersebut diantaranya *preprocessing*, *tokenization,* menghilangkan kata-kata yang bukan bahasa arab, menghilangkan nomor, karakter spesial, tanda baca, mengihlangkan kata awalan dan akhiran serta proses *stemming*. Setelah itu tahap pengklasifikasian menggunakan regresi logistik, *Support Vector Machine, Naive Baiyes* dan GIS. Dari penelitian, terbukti *Support Vector Machines* memiliki tingkat akurasi lebih tinggi dibandingkan dengan pengklasifikasi lainnya dalam hal ini regresi logistik, *Naive Baiyes* dan GIS (Al-tahrawi, 2015).

2.1.6. Metode Penelitian Chou

Menurut Chou (Chou et al., 2014) dalam penelitiannya membahas mengenai bagaimana mengoptimalkan parameter *Support Vector Machine* menggunakan algoritma genetika untuk klasifikasi kecenderungan sengketa di tahap awal proyek kemitraan publik-swasta. Dataset yang digunakan dalam penelitian ini berdasrkan data yang diambil dari *Taiwan Public Construction Commission* (TPCC) – otorisasi pemerintah yang mengawasi pelayanan publik dan infrastruktur kontruksi di Taiwan. *Database* penelitian berisi 584 proyek PPP diawasi oleh TPCC sejak tahun 2002 sampai tahun 2009.

Studi ini mengusulkan model intelejen yang dioptimalkan untuk mengintegrasikan *Fast Messy Genetic Algorithms (*fmGA) dengan *Support Vector Machine* (SVM). SVM berbasis fmGA (GASVM) digunakan untuk prediksi awal dari kecenderungan sengketa di tahap awal proyek kemitraan publik-swasta. SVM menyediakan pembelajaran dan *curve fitting* sedangkan fmGA mengoptimalkan parameter SVM. Langkah-langkah yang terjadi seperti akurasi, presisi, sensitivitas, spesifitas dan area di bawah kurva dan indeks sintesis yang digunakan untuk evaluasi kinerja yang diusulkan. Ini menunjukan bahwa GASVM mencapai akurasi prediksi dalam *Fold Cross Validation* lebih baik dibandingkan dengan model lainnya (yaitu, CART, CHAID, QUEST, dan C5.0). dimana GASVM mencapai akurasi sebesar 89.30%, CART 80%, CHAID 82.63%, QUEST 79.06% dan C 5.0 83.25%.

Dari penelitian tersebut, GA dan *Support Vector Machine* (SVM) memiliki tingkat akurasi lebih tinggi daripada pengklasifikasi lainnya (Chou et al., 2014).

2.2. Tinjauan Pustaka

2.2.1. *Text Mining*

Text mining adalah salah satu bidang khusus dari *data mining*. *Text mining* dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang pengguna berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis, yang merupakan komponen-komponen dalam *data mining* salah satunya adalah klasifikasi. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Maka dari itu, sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks *(text categorization)* dan pengelompokan teks *(text clustering)* (Feldman dan Sanger, 2007).

Text mining bisa dianggap subjek riset yang tergolong baru. *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian / pengelompokkan dan menganalisa *unstructured text* dalam jumlah besar. Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *Data mining*, *Information Retrieval*, Statistik dan Matematik, *Machine Learning*, *Linguistic*, *Natural Languange Processing*, dan *Visualization*. Kegiatan riset untuk *text mining* antara lain ekstraksi dan penyimpanan teks, *preprocessing* konten teks, pengumpulan data statistik dan *indexing*, dan analisa konten. Gambar 2.1. merupakan gambaran umum tentang kerangka text mining.

Gambar 2.1. *High-Level Text Mining Functional Architecture*

Dari gambar diatas menjelaskan alur atau proses dari pengolahan dokumen, dimana dimulai dari pengumpulan data, lalu dilakukan proses *preprocessing* yang didalamnya terdapat proses *categorization, feature/term extraction,* setelah melalui proses tersebut dokumen masuk kedalam proses pemberian *keyword,labeled dan time-stamped,* dan terakhir adalah proses *core mining operations* dan *presentation.*

2.2.2. Klasifikasi Teks (*Classification Text)*

Klasifikasi teks merupakan bagian penting dari text mining. Klasifikasi teks berdasarkan para ahli, bagaimana

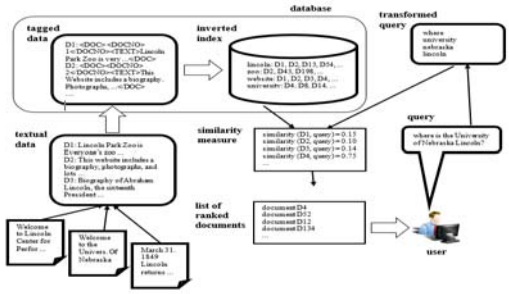
mengklasifikasikan dokumen didalam beberapa kategori-kategori (Kaur, 2013).

Klasifikasi atau kategorisasi teks adalah proses penempatan suatu dokumen ke suatu kategori atau kelas sesuai dengan karakteristik dari dokumen tersebut. Dalam *text mining*, klasifikasi mengacu kepada aktifitas menganalisis atau mempelajari himpunan dokumen teks *preclasified* untuk memperoleh suatu model atau fungsi yang dapat digunakan untuk mengelompokkan dokumen teks lain yang belum diketahui kelasnya ke dalam satu atau lebih kelas *pre-defined* (Sebastiani, 2002).

2.2.3. *Information Retrieval*

ISO 2382/1 standard, mendefinisikan *Information Retrieval* sebagai tindakan, metode dan prosedur untuk menemukan kembali data yang tersimpan, kemudian menyediakan informasi mengenai subjek yang dibutuhkan. Tindakan tersebut mencakup *text indexing, inquiry analysis,* dan *relevance analysis.* Data mencakup teks, tabel, gambar, ucapan, dan video. Informasi termasuk pengetahuan terkait yang dibutuhkan untuk mendukung penyelesaian masalah dan akuisisi pengetahuan (Cios et al., 2007).

Tujuan dari sistem *Information Retrieval* adalah untuk memenuhi kebutuhan informasi pengguna dan me-*retreive* semua dokumen yang mungkin relevan, pada waktu bersamaan me-*retreive* sedikit mungkin dokumen yang tidak relevan. Sistem ini menggunakan fungsi heuristik untuk mendapatkan dokumen-dokumen yang relevan dengan *query* pengguna. Sistem *Information Retrieval* yang baik memungkinkan pengguna menentukan secara cepat dan akurat apakah isi dari dokumen yang diterima memenuhi kebutuhannya. Agar representasi dokumen lebih baik, dokumen-dokumen dengan topik atau isi yang mirip dikelompokkan bersama-sama (Azrifah dan Murad, 2007).



Gambar 2.2. Arsitektur Dasar Sistem *Information Retrieval*

2.2.4. Fase Klasifikasi Utama

A. *Document Collection*

Kunci utama dari *text mining* adalah koleksi dokumen (*document collection*). Koleksi dokumen dapat berupa kumpulan dari dokumen yang berbasis teks. Kumpulan dokumen tersebut dapat terus bertambah dari ratusan, ribuan, bahkan bisa menjadi jutaan dokumen. Dari pertumbuhan dokumen tersebut, *text mining* mempunyai tugas penting, yaitu menemukan pola dari jumlah dokumen yang begitu besar.

B. *Preprocessing*

Pembangunan *index* dari kolesi dokumen merupakan tugas pokok pada tahapan *preprocessing* di dalam *information retrieval*. Kualitas *index* mempengaruhi efektifitas dan efesiensi sistem IR. *Index* dokumen adalah himpunan *term* yang menunjukkan isi atau topik yang dikandung oleh dokumen. *Index* akan membedakan suatu dokumen dari dokumen lain yang berada di dalam koleksi. Ukuran *index* yang kecil dapat memberikan hasil buruk dan mungkin beberapa *term* yang relevan terabaikan. *Index* yang besar memungkinkan ditemukanya banyak dokumen yang relevan tetapi sekaligius dapat menaikkan jumlah dokumen yang tidak relevan dan menurunkan kecepatan pencarian (*searching*). Pembuatan *inverted index* harus melibatkan konsep *linguistic processing* yang bertujuan mengesktrak *term-term* penting dari dokumen yang direpresentasikan sebagai *bag-of-words*. Ekstraksi *term* biasanya melibatkan dua operasi utama yaitu penghapusan *stop-words* dan *stemming* (Cios et al., 2007). Sedangkan *indexing* menerapkan beberapa langkah diantaranya penghapusan format dan *markup* dari dalam dokumen, *tokenization*, penyaringan (*Filter*), *Stemming, p*emberian bobot terhadap *term* (*weighting*) (Manning, Ragnavan, dan Schutze, 2008).

2.2.5. Algoritma *Support Vector Machines* (SVM)

Support Vector Machines (SVM) adalah suatu metode klasifikasi pola populer dengan banyak aplikasi yang beragam. Pengaturan parameter kernel dalam prosedur pelatihan SVM, bersama dengan pilihan fitur, secara signifikan mempengaruhi akurasi klasifikasi (Lin et al., 2008).

Support Vector Machines (SVM) adalah seperangkat metode yang terkait untuk suatu metode pembelajaran, untuk kedua masalah klasifikasi dan regresi (Maimon, 2010).

Secara konseptual, SVM adalah mesin linear, dilengkapi dengan fitur-fitur khusus, dan berdasarkan minimisasi risiko struktural (SRM) metode dan teori belajar statistik (Gorunescu, 2011).

Sedangkan menurut (Berry dan Cogan, 2010) SVM adalah model linier untuk menerapkan batas-batas kategori nonlinier dengan mengubah ruang contoh yang diberikan

menjadi linear dipisahkan satu sampai pemetaan nonlinear.
Menurut Lee (Lee dan Yang, 2009) teknik *One-Against-All* memungkinkan penyebaran pengklasifikasi kebih sedikit untuk mencapai fungsi kategorisasi *multiclass* dan juga mendapatkan kinerja yang wajar dalam sebuah aplikasi. Berikut contoh ilustrasi metode *One-Against-All:*

Tabel 2.1 Ilustrasi Metode One-againts-all

$y_i = 1$	$y_i = -1$
Kelas 1	Bukan kelas 1
Kelas 2	Bukan kelas 2
Kelas 3	Bukan kelas 3
Kelas 4	Bukan kelas 4

2.2.6. *Particle Swarm Optimization* (PSO)

Particle Swarm Optimization (PSO) pada awalnya dirancang dan diperkenalkan oleh Eberhart dan Kennedy. PSO adalah algoritma pencarian penduduk berdasarkan berdasarkan simulasi perilaku sosial burung, lebah atau sekolah ikan. Algoritma ini awalnya bermaksud untuk grafis mensimulasikan koreografi anggun dan tak terduga dari rakyat burung .

Particle Swarm Optimization (PSO) adalah suatu teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter (Basari et al., 2013).
Particle Swarm Optimization (PSO) adalah berdasarkan populasi meta-heuristik baru yang mensimulasikan perilaku sosial seperti burung berbondong-bondong ke posisi menjanjikan untuk mencapai tujuan yang tepat dalam ruang multidimensi (Lin et al., 2008).

2.3. Kerangka Pemikiran

Penelitian ini dimulai dari adanya masalah dalam klasifikasi teks pada konten berita digital Bahasa Indonesia menggunakan pengklasifikasi *Support Vector Machines* (SVM), dimana 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 konten berita digital bahasa Indonesia yang didapat dari www.kompas.com yang terdiri dari 80 konten berita gosip, 80 konten berita travel dan, 80 konten berita kuliner. *Preprocessing* yang dilakukkan dengan *Tokenization, Tranform Cases.* Metode pembobotan Fitur yang akan digunakan adalah *Term Frequency Invers Documet Frequency* (TF-IDF) dan pemilihan seleksi fitur menggunakan *Particle Swarm Optimization* (PSO). Sedangkan pengklasifikasi yang digunakan adalah *Support Vector Machines*. Pengujian 10 *Fold Cross Validation* akan dilakukan dan akurasi algoritma akan diukur menggunakan *Confusion Matrix*. *RapidMiner* Versi 5.3 digunakan sebagai alat bantu dalam mengukur akurasi data eksperimen yang dilakukkan dalam penelitian.

3.1. Perancangan Penelitian

Metode penelitian yang penulis lakukan adalah metode penelitian eksperimen, dengan tahapan sebagai berikut:
1. Pengumpulan Data
Data untuk eksperimen ini dikumpulkan, lalu diseleksi dari data yang tidak sesuai.
2. Pengolahan Awal Data
Model dipilih berdasarkan kesesuaian data dengan metode yang paling baik dari beberapa metode pengklasifikasian teks yang sudah digunakan oleh beberapa peneliti sebelumnya. Model yang digunakan adalah algoritma *Support Vector Machines*.
3. Metode Yang Diusulkan
Untuk meningkatkan akurasi dari Algoritma *Support Vector Machines*, maka dilakukan penambahan dengan menggabungkan metode peningkatan optimasi yaitu *Particle Swarm Optimization* (PSO).
4. Eksperimen dan Pengujian Metode
Untuk eksperimen data penelitian, penulis menggunakan *RapidMiner* 5.3 untuk mengolah data dan sebagai alat bantu dalam mengukur akurasi data eksperimen yang dilakukan dalam penelitian.
5. Evaluasi dan Validasi Hasil
Evaluasi dilakukan untuk mengetahui akurasi dari model algoritma *Support Vector Machines*. Validasi digunakan untuk melihat perbandingan hasil akurasi dari model yang digunakan dengan hasil yang telah ada sebelumnya. Teknik validasi yang digunakan adalah *Cross Validation*, akurasi algoritma akan diukur menggunakan *Confusion Matrix*.

3.2. Pengumpulan Data

Penulis menggunakan data konten berita digital Bahasa Indonesia yang dikumpulkan dari situs www.kompas.com Data terdiri dari 80 konten berita gosip, 80 konten berita travel, dan 80 konten berita kuliner.

3.3. Pengolahan Awal Data

Untuk mengurangi lamanya waktu pengolahan data, penulis hanya menggunakan 80 konten berita gosip, 80 konten berita travel, dan 80 konten berita kuliner sebagai data training. *Dataset* ini dalam tahap *preprocessing* harus melalui dua proses, yaitu:
1. *Tokenization*

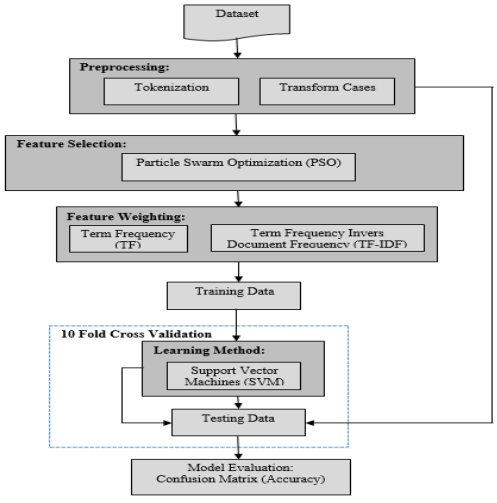
Yaitu mengumpulkan semua kata yang muncul dan menghilangkan tanda baca maupun simbol apapun yang bukan huruf.

2. *Tranform Cases*
   Yaitu merubah semua huruf besar/*uppercase* menjadi huruf kecil/*lowercase* sehingga semua huruf yang diproses berjenis huruf kecil/*lowercase*

Sedangkan untuk tahap *transformation* dengan melakukan pembobotan TF-IDF pada masing-masing kata. Di mana prosesnya menghitung kehadiran atau ketidakhadiran sebuah kata di dalam dokumen. Berapa kali sebuah kata muncul didalam suatu dokumen juga digunakan sebagai skema pembobotan dari data tekstual.

3.4. Metode Yang Diusulkan

Metode yang penulis usulkan adalah penggunaan satu jenis metode pemilihan fitur, yaitu *Particle Swarm Optimization* (PSO) yang digunakan sebagai metode pemilihan fitur agar akurasi pengklasifikasi *Support Vector Machines* (SVM) bisa meningkat. Penulis menggunakan pengklasifikasi *Support Vector Machines* karena merupakan teknik *machine learning* yang populer untuk klasifikasi teks, serta memiliki performa yang baik pada banyak domain. SVM memiliki kelebihan yaitu mampu mengidentifikasi *hyperplane* terpisah yang memaksimalkan margin antara dua kela yang berbeda. SVM menjamin untuk memaksimalkan jarak antara data yang paling dekat dengan *hyerplane.* Jika masukan data dapat dipisahkan secara linier. Lihat gambar III.1. untuk model yang diusulkan secara lebih detail.



Gambar 3.1. Model Yang Diusulkan

Hasil yang dibandingkan adalah akurasi *Support Vector Machines* (SVM) sebelum menggunakan metode pemilihan fitur dengan akurasi *Support Vector Machines* (SVM) setelah menggunakan metode pemilihan fitur yaitu *Particle Swarm Optimization* (PSO).

3.5. Eksperimen dan Pengujian Model

Penulis melakukkan proses eksperimen menggunakan aplikasi *RapidMiner* 5.3. sedangkan untuk pengujian model dilakukkan menggunakan *dataset* konten berita digital bahasa Indonesia dari situs www.kompas.com yang telah dikategorikan ke dalam konten berita gosip, konten berita travel, dan konten berita kuliner. Sedangkan untuk pengujian model dilakukkan menggunakan *dataset* konten berita digital bahasa Indonesia yang berbeda dari data training. Sedangkan spesifikasi komputer yang penulis gunakan dalam penelitian ini dapat dilihat pada

Tabel 3.1. Spesifikasi minimum komputer yang digunakan

Processor	*Intel Core* i3 2.20 GHz
Memori	2 GB
Harddisk	250 GB
Sistem Operasi	*Microsoft Windows* 7
Aplikasi *Text Mining*	*RapidMiner* versi 5.3
Perangkat Lunak	*Adobe Dreamweaver* CS 5
Bahasa Pemrograman	PHP

3.6. Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang klasifikasi konten berita digital bahasa Indonesia adalah dengan menerapkan *Support Vector Machines* (SVM) dan *Support Vector Machines* (SVM) berbasis *Particle Swarm Optimization* (PSO). Penerapan algoritma SVM dengan menentukan tipe kernel lebih dahulu. Setelah didapatkan nilai akurasi. Sedangkan penerapan algoritma SVM berbasis PSO beracuan pada penentuan nilai *population size* yang tepat. Dari nilai akurasi yang paling ideal dari parameter tersebut, terbentuk struktur algoritma yang ideal untuk pemecahan masalah tersebut.

4.1. Hasil

4.1.1. Klasifikasi Teks Menggunakan Algoritma *Support Vector Machine*

Data training yang digunakan dalam pengklasifikasian teks ini terdiri dari 80 konten berita gosip, 80 konten berita travel, dan 80 konten berita kuliner. Data tersebut masih berupa sekumpulan teks yang terpisah dalam bentuk dokumen. Sebelum diklasifikasikan, data tersebut harus melalui beberapa tahapan proses agar bisa diklasifikasikan dalam proses selanjutnya, berikut adalah tahapan prosesnya:

1. Pengumpulan Data
   Data berita gosip disatukan dalam folder dengan nama gosip, Data berita travel disatukan dalam folder dengan nama travel, sedangkan data berita kuliner disatukan dalam folder dengan nama

kuliner. Tiap dokumen berekstensi .txt yang dapat dibuka menggunakan aplikasi *Notepad.*

2. Pengolahan Awal Data

Proses yang dilalui terdiri dari *tokenization, stopwords removal,* dan *stemming.*

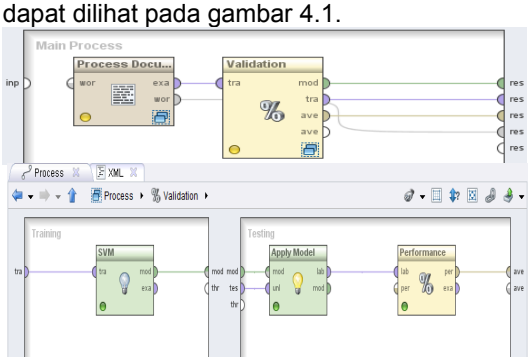
Tabel 4.1. Pengolahan Awal Data

Konten	*Tokenization*	*Transform Cases*
Tim basket putra Indonesia gagal menyumbang emas di ajang SEA Games 2015. Tim 'Merah-Putih' cuma menyumbang perak setelah takluk 64-72 dari Filipina di babak final.	Tim basket putra Indonesia gagal menyumbang emas di ajang SEA Games 2015 Tim Merah Putih cuma menyumbang perak setelah takluk 6472 dari Filipina di babak final	tim basket putra indonesia gagal menyumbang emas di ajang sea games 2015 tim merah putih cuma menyumbang perak setelah takluk 6472 dari filipina di babak final

3. Klasifikasi

Proses klasifikasi disini adalah untuk menentukan sebuah kalimat sebagai anggota *class* gosip, *class* travel, dan *class* kuliner berdasarkan nilai perhitungan probabilitas dari rumus SVM yang lebih besar. Jika hasil probabilitas kalimat tersebut untuk *class* gosip lebih besar daripada *class* travel dan kuliner, maka kalimat tersebut termasuk dalam *class* gosip. Begitu juga sebaliknya dengan *class* travel, dan kuliner. Penulis mengambil dikumen keseluruhan sebanyak 240 data *training* dan 5 kata yang berhubungan dengan masing-masing konten berita, berikut kata yang berhubungan dengan konten berita gosip yaitu gosip, selebriti, selingkuh, artis, skandal. Lalu berikut kata yang berhubungan dengan konten berita travel yaitu wisata, pantai, travel, trip, gunug. Sedangkan kata yang berhubungan dengan konten kuliner yaitu makan, minum, restoran, lezat, kuliner.

Penulis membuat model dengan menggunakan RapidMiner 5. Desain model dapat dilihat pada gambar 4.1.



Gambar 4.1. Desain model Support Vector Machines menggunakan RapidMiner

4.1.2. Pengujian Model dengan 10 *Fold Cross Validation*

Pada penelitian ini, penulis melakukan pengujian model dengan menggunakan teknik 10 *cross validation,* di mana proses ini membagi data secara acak ke dalam 10 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya. Setelah itu proses akurasi dihitung dengan melihat seberapa banyak data yang sudah terklasifikasi dengan benar.

Tabel 4.2 Pengujian 10 Fold Cross Validation

Cross Validation	SVM + PSO *Accuracy*
2	94.50 %
3	65.04 %
4	94.25 %
5	94.25 %
6	32.08 %
7	68.05 %
8	47.75 %
9	69.23 %
10	95.42 %

4.2. Evaluasi dan Validasi Hasil

Hasil dari pengujian model yang dilakukan adalah mengklasifikasikan berita gosip, berita travel, dan berita kuliner dari suatu konten berita dengan *Support Vector Machine* dan *Support Vector Machine* berbasis *Particle Swarm Optimization* untuk menentukan nilai *accuracy*. Dalam menentukan nilai tingkat keakurasian dalam model *Support Vector Machine* dan algoritma *Particle Swarm Optimization.*

4.2.1. Analisis Evaluasi Hasil dan Validasi Model

Dari hasil di atas, pengukuran akurasi menggunkan *confusion matrix* terbukti bahwa hasil pengujian algoritma *Support Vector Machine* berbasis *Particle Swarm Optimization* memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma SVM. Nilai akurasi untk model algoritma SVM sebesar 65.81% dan nilai akurasi unutk model algoritma SVM berbasis PSO sebesar 95.42% dengan selisih akurasi 29.61%.

4.3. Pembahasan

Berdasarkan hasil eksperimen yang dilakukan untuk memecahkan masalah klasifikasi konten berita digital, dapat disimpulkan bahwa hasil eksperimen menggunakan metode *Support Vector Machine* mempunyai tingkat akurasi sebesar 65.81% .

Sedangkan eksperimen kedua yang dilakukan dengan menggunakan metode *Support Vector Machine* berbasis *Particle Swarm Optimization* didapat nilai akurasi terbaik yaitu mempunyai akurasi sebesar 95.42%.

Tabel 4.3 Model *Confusion Matrix* Untuk Metode *Support Vector Machine*

Accuracy: 65.81% +/- 31.63% (mikro: 66.25%)				
	true Travel	true Gosip	true Kuliner	class precision
Pred Travel	53	18	18	59.55%
Pred Gosip	14	53	9	69.74%
Pred Kuliner	13	9	53	70.67%
Class recall	66.25 %	66.25%	66.25%	

Tabel 4.4 Model *Confusion Matrix* Untuk Metode *Support Vector Machine* Berbasis *Particle Swarm Optimization*

Accuracy: 95.42% +/- 2.24% (mikro: 95.42%)				
	true Travel	true Gosip	true Kuliner	class precision
Pred Travel	75	0	4	94.94 %
Pred Gosip	3	80	2	94.12 %
Pred Kuliner	2	0	74	97.37 %
Class recall	93.75%	100%	92.50%	

4.4. Implikasi Penelitian

Implikasi penelitian ini mencakup beberapa aspek, diantaranya:

1. Implikasi terhadap aspek sistem
   Hasil evaluasi menunjukkan penerapan *Particle Swarm Optimization* untuk seleksi fitur dapat meningkatkan akurasi *Support Vector Machine* dan merupakan metode yang cukup baik dalam mengklasifikasi konten berita digital. Dengan demikian penerapan metode tersebut dapat membantu para penyedia berita digital bisa membuat pekerjaan lebih efektif dan efisien mungkin.
2. Implikasi terhadap aspek manajerial
   Membantu para pengembang sistem yang berkaitan dengan perusahaan konten berita digital, maupun media sosial dan lain-lain agar menggunakan aplikasi RapidMiner dalam membangun suatu sistem.
3. Implikasi terhadap aspek penelitian lanjutan
   Penelitian selanjutnya bisa menggunakan metode pemilihan fitur ataupun *dataset* dengan bahasa dan dari domain yang berbeda, seperti konten berita bahasa Yunani, konten berita bahasa China, konten berita bahasa Spanyol dan sebagainya.

5. KESIMPULAN

Dalam penelitian ini dilakukan pengujian model dengan menggunakan *Support Vector Machine* dan *Support Vector Machine* berbasis *Particel Swarm Optimization* dengan menggunakan data konten berita dengan keseluruhan 240 data konten berita dan 15 kata yang berhubungan dengan konten berita tersebut, yaitu gosip, selebriti, selingkuh, artis, skandal wisata, pantai, travel, trip, gunug makan, minum, restoran, lezat, kuliner. Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy, precision,* dan *recall* dari setiap algoritma sehingga didapatkan pengujian dengan menggunakan *Support Vector Machine* didapatkan nilai *accuracy* adalah 65.81%. Sedangkan pengujian dengan menggunakan *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO) didapatkan nilai *accuracy* 95.42%. Maka dapat disimpulkan pengujian data konten berita digital menggunakan *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO) lebih baik dari pada *Support Vector Machine* sendiri. Dengan demikian hasil dari pengujian model di atas dapat disimpulkan bahwa *Support Vector Machine* berbasis *Particle Swarm Optimization* (PSO) memberikan pemecahan untuk permasalahan klasifikasi konten berita digital lebih akurat.

Walaupun pengklasifikasi *Support Vector Machines* sudah sering digunakan dan mempunyai performa yang baik dalam mengklasifikasikan teks, namun ada beberapa hal yang dapat ditambahkan untuk penelitian selanjutnya:

1. Menggunakan metode pemilihan fitur yang lain, seperti *Chi Square, Gini Index, Mutual Information, Genetic Algorithm* dan lain-lain agar hasilnya bisa dibandingkan.
2. Menggunakan pengklasifikasi lain yang mungkin di luar *Supervised learning.* Sehingga bisa dilakukan penelitian yang berbeda dari umumnya yang suda ada.
3. Menggunakan data konten berita yang berbeda dari bahasa yang berbeda, misalnya konten berita bahasa Yunani, konten berita bahasa China, konten berita bahasa Spanyol dan sebagainya.

DAFTAR PUSTAKA

[1] Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *Int. Arab J. E-Technol.*, *2*(2), 124–128.

[2] Al-tahrawi, M. M. (2015). Arabic Text Categorization Using Logistic Regression, (May), 71–78.

[3] Azrifah, M., & Murad, A. (2007). Word Similarity for Document Grouping using Soft Computing, *7*(8), 20–28.
[4] Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, *53*, 453–462.
[5] Chou, J. S., Cheng, M. Y., Wu, Y. W., & Pham, A. D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. *Expert Systems with Applications*, *41*(8), 3955–3964.
[6] Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). Data Mining
[7] A Knowledge Discovery Approach. Springer.
[8] Feldman, R., & Sanger, J. (2007). The Text Mining Handbook Advanced
[9] Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press.
[10] Gorunescu, F. (2011). *Data Mining Concept Model Technique*.
[11] Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Soft Computing* (Vol. 54).
[12] Haupt, S. E. (2004). *Practical Genetic Algorithms*.
[13] Janpla, S. (2014). The Effectiveness Of Automated Thai Documents Categorization Based On Machine, *66*(1).
[14] Kaizhu Huang, Haiqin Yang, Irwin King, M. L. (2008). *Advanced Topics in Science and Technology in China*.
[15] Kaur, H. (2013). Online News Classification□: A Review, 7–9.
[16] Krishnalal, G., Rengarajan, S. B., & Srinivasagan, K. G. (2010). A New Text Mining Approach Based on HMM-SVM for Web News Classification. *International Journal of Computer Applications*, *1*(19), 103–109.
[17] Lee, C.-H., & Yang, H.-C. (2009). Construction of supervised and unsupervised learning systems for multilingual text categorization. *Expert Systems with Applications*, *36*(2), 2400–2410.
[18] Lin, S.-W., Ying, K.-C., Chen, S.-C., & Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, *35*(4), 1817–1824.
[19] Liu, Y., Wang, G., Chen, H., & Dong, H. (2011). An improved particle swarm optimization for feature selection. *Journal of Bionic …*, *8*, 392–397.
[20] Mahinovs, A. Tiwarigton, a. (2007). Text Classification Method Review. *Decision Engineering Report Series*, (April).
[21] Maimon, O. (2010). *Data Mining And Knowledge Discovery Handbook.* New York Dordrecht Heidelberg London: Springer.
[22] Manning, C. D., Ragnavan, P., & Schutze, H. (2008). An Introduction to Information Retreival. *IEEE Photonics Technology Letters*, *21*(8), C3–C3.
[23] Polettini, N. (2004). The Vector Space Model in Information Retrieval - Term Weighting Problem. 1-9.
[24] Sebastiani, F. (2001). Machine Learning in Automated Text Categorization.
[25] W.Berry, Michael(University of Tennessee, U., & Kogan, Jacob(University of Maryland Baltimore Country, U. (2010). *Text Mining Application and Theo*