Implementasi Algoritma Naïve Bayes Classifier Berbasis Particle Swarm Optimization (PSO) Untuk Klasifikasi Konten Berita Digital Bahasa Indonesia

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

Abstract - A lot of important information is stored in the document word, and have each topic, then text classification is one solution to manage the information that is growing rapidly and the abundant, and already many agencies engaged in the distribution of information or news already started using web-based systems to deliver up to date news. However, the news divide into these categories for now still dilakukkan manually, so it is very troublesome and can also take a long time. In this study will be used merging feature selection methods, namely Particle Swarm Optimization based Naïve Bayes classifier to look at the accuracy of the method. This research has resulted in the form of text classification category of gossip, culinary, and travel from digital news content. Measurement is based on Naïve Bayes classifier accuracy before and after the addition of feature selection methods. The evaluation was done using a 10 fold cross validation. While the measurement accuracy is measured by confusion matrix. The results of this study obtained accuracy by using Naïve Bayes classifier algorithm method amounted to 94.17%.
Keywords: Particle Swarm Optimization, Naïve Bayes classifier, classification News Content, Text Mining

Abstrak - Banyak informasi penting yang tersimpan didalam dokumen berita, dan mempunyai topik masing-masing, kemudian klasifikasi teks merupakan salah satu solusi untuk mengelola informasi yang berkembang pesat dan melimpah tersebut, serta sudah banyak juga 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. Dalam penelitian ini akan digunakan penggabungan metode pemilihan fitur, yaitu *Particle Swarm Optimization* berbasis *Naïve Bayes Classifier* untuk melihat akurasi pada metode tersebut*.* Penelitian ini menghasilkan klasifikasi teks dalam bentuk kategori gosip, kuliner, dan travel dari konten berita digital. Pengukuran berdasarkan akurasi *Naïve Bayes Classifier* sebelum dan sesudah penambahan metode pemilihan fitur. Evaluasi dilakukan menggunakan 10 *fold cross validation.* Sedangkan pengukuran akurasi diukur dengan *confusion matrix.* Hasil penelitian ini didapat akurasi dengan menggunakan metode algoritma *Naïve Bayes Classifier* sebesar 94.17%.
Kata kunci : Particle Swarm Optimization, Naïve Bayes Classifier, Klasifikasi Konten Berita, Text Mining

1. PENDAHULUAN

Kehadiran komputer personal membuat proses digitalisasi sebuah informasi menjadi lebih mudah, hanya dengan sebuah perangkat lunak pengolah kata, setiap orang dapat menulis setiap kejadian langsung dalam format digital. 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 www.kompas.com yang merupakan salah satu *website* berita yang sering dikunjungi. Maka dari itu banyak dari media informasi dimasa sekarang yang melakukan pengklasifikasian banyak informasi yang dapat kita terima dalam *website* tersebut. Terkadang, kita langsung saja menerima tanpa adanya penyeleksian informasi. Atas dengan kategorisasi terlebih dahulu sebelum disebarkan pada masyarakat luas. Pengklasifikasian tersebut berguna untuk memudahkan masyarakat untuk mencari informasi yang mereka inginkan.

Penggunaan situs berita *online* menjadi cara baru yang banyak digunakan saat ini. Cara konvensional seperti koran harian, majalah mingguan atau bulanan masih digunakan untuk menjangkau khalayak yang masih menyukai bentuk konvensional. Penggunaan situs dengan *world wide web* menghadirkan fitur *hyperlink* yang memberikan kemudahan kepada pembaca untuk menelusuri informasi-informasi lanjutan mengenai topik dalam sebuah berita di lokasi yang berbeda. Penggunaan situs memungkinkan penyebaran lebih cepat, aktual, lebih murah, dan ramah lingkungan.

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, Klasifikasi dokumen teks berbahasa Indonesia menggunakan Naïve Bayes (Samodra, Sumpeno, & Hariadi, 2009), Klasifikasi dokumen berbahasa Indonesia menggunakan Klasifikasi berita berbahasa Indonesia menggunakan Naïve Bayes Classifier (Efendi, Malik, dan Sari, 2012), Klasifikasi berita lokal radar Malang menggunakan metode Klasifikasi berita berbahasa Indonesia menggunakan *Naïve Bayes* dengan fitur N-Gram (Chandra, Indrawan, dn Sukajaya, 2016).

Dari beberapa penelitian tersebut, teknik yang banyak digunakan untuk klasifikasi data adalah *Naïve Bayes Classifier*.

Ada beberapa teknik yang bisa digunakan untuk lebih menyempurnakan kekurangan *Naïve Bayes Classifier* 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 Naïve Bayes Classifier 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. LANDASAN/KERANGKA PEMIKIRAN

2.1. Tinjauan Studi

Ada beberapa penelitian yang menggunakan *Naïve Bayes Classifier* sebagai pengklasifikasi dalam klasifikasi teks isi konten berita digital, diantaranya:

2.1.1. Model Penelitian Joko

Penelitian yang dilakukkan oleh Joko Samodra, Surya Sumpeno, dan Mochamad Hariadi mengenai klasifikasi dokumen teks berbahasa Indonesia dengan menggunakan Naïve Bayes Classifier. Data yang digunakan sebagai sampel dalam penelitian ini diambil dari www.tempointeraktif.com edisi tanggal 1 April 2002 sampai 30 Juni 2002, dimana jumlah dokumen yang digunakan adalah 2.400 dokumen, yang terbagi menjadi 4 kategori yaitu: Nasional, Metro, Nusantara, dan Ekonomi Bisnis. Langkah-langkah yang dilakukan pada tahap awal (*preprocessing),* dokumen yang telah didownload dan berbentuk file html diedit secara manal untuk diambil bagian teksnya saja. Setelah selesai diedit dokumen tersebut dipisahkan menjadi 4 bagian sesuai dengan kategorinya masing-masing. Sedangkan program yang digunakan untuk melakukan percobaan adalah MALLET 2.0.

Dari hasil penelitian ini menunjukkan bahwa metode *Naïve Bayes Classifier* dapat digunakan secara efektif untuk mengklasifikasi dokumen teks berbahasa Indonesia. Hal ini terlihat dari hasil eksperimen yaitu dengan porsi dokumen training yang kecil (20%) nilai akurasinya dapat mencapai 83,57%, dan terus meningkat hingga 87,63% sesuai dengan peningkatan porsi dokumen training (Samodra, Sumpeno, dan Hariadi, 2009).

2.1.2. Model Penelitian Efendi

Penelitian yang dilakukkan oleh Rusdi Efendi, Reza Firsandaya Malik, dan Jeni Mila Sari U mengenai sebuah klasifikasi dokumen berbahasa Indonesia menggunakan Klasifikasi berita berbahasa Indonesia menggunakan *Naïve Bayes Classifier*. Dimana *dataset* dalam pengujian ini menggunakan 60 dokumen berita dengan berbagai kategori dari media elektronik yaitu, ekonomi, kesehatan, olahraga, teknologi, politik dan pendidikan. Pada setiap proses klasifikasi ataupun pelatihan, semua dokumen yang digunakan harus melewati proses *text mining* terlebih dahulu, yaitu proses *tokenizing* (pemecahan kata), *filtering* (penyaringan kata), *stemming* (penghilangan imbuhan). Pada proses pelatihan yang dilakukan terbentuklah token atau kosakata.

Dari hasil penelitian tersebut, keakuratan yang didapat oleh aplikasi ini mencapai 86,67%, dimana jumlah pengujian 30 contoh dokumen uji dan hanya 26 contoh dokumen uji yang memiliki nilai benar. Hal ini dikarenakan dokumen pelatihan yang sedikit menyebabkan kurangnya kata-kata yang penting yang mencirikan suatu dokumen dan juga terdapat kata-kata yang dominan ke kategori lain yang bukan kategorinya sehingga dapat menimbulkan kesalahan dalam pengklasifikasian dokumen (Efendi, Malik, dan Sari, 2012).

2.1.3. Model Penelitian Chandra

Penelitian yang dilakukan oleh Denny Nathaniel Chandra, Gede Indrawan, dan I Nyoman Sukajaya mengenai klasifikasi berita

lokal radar malang menggunakan metode Klasifikasi berita berbahasa Indonesia menggunakan *Naïve Bayes* dengan fitur N-Gram, tahapan yang ada dalam klasifikasi meliputi *converting* dan *filtering* kemudian pemrosesan file seperti proses pengenalan pola klasifikasi, metode pengukuran dan hasil pengukuran, kualitas informasi pada klasifikasi dokumen menggunakan metode *Naïve Bayes Classification.* Metode *Naïve Bayes Classification* menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin merepresentasikan dokumen, pada tahap pelatihan terdapat dokumen *training* yang menjadi acuan untuk proses *testing*. Dimana dalam tahapan tersebut terdapat proses *preprocessing* dari dokumen *training* maupun *testing,* setelah itu proses tokenisasi lalu proses *stopword.*

Dari penelitian tersebut didapatkan hasil akurasi setelah melakukan 5 kali percobaan, akurasi yang didapat sebesar 78,66% yang didapat dari percobaan pertama (Chandra, Indrawan, dan Sukajaya, 2016).

2.2. Tinjauan Pustaka

2.2.1. *Text Mining*

Penambangan teks (bahasa inggris: *text mining*) adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen Word, PDF, kutipanteks, dll. Jenis masukan untuk penambangan teks ini disebut data tidak terstruktur dan merupakan pembeda utama dengan penambangan data yang menggunakan data terstruktur atau basis data sebagai masukan (Chandra, Indrawan, dan Sukajaya, 2016).

Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevansi data teks terstruktur ini dengan menggunakan teknik dan alat yang sama dengan penambangan data. Proses yang umum dilakukan oleh penambangan teks diantaranya adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks, dan lain-lain.

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).

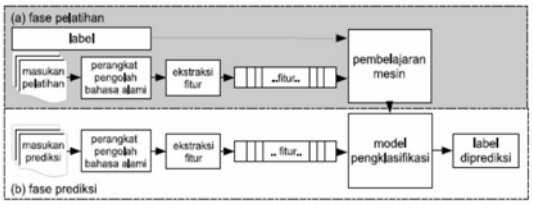
2.2.2. Berita

Berita merupakan bentuk laporan tentang suatu kejadian yang sedang terjadi baru-baru ini atau keterangan terbaru dari suatu peristiwa. Dengan kata lain berita adalah fakta menarik atau sesuatu hal yang penting yang disampaikan pada masyarakat orang banyak melalui media. Tapi tidak semua fakta bisa diangkat menjadi suatu berita oleh media. Karena setiap fakta akan dipilih mana yang pantas untuk disampaikan pada masyarakat.

2.2.3. Klasifikasi Teks (*Classification Text)*

Pengklasifikasian teks sangat dibutuhkan dalam berbagai macam aplikasi, terutama aplikasi yang jumlah dokumennya bertambah dengan cepat. Ada dua cara dalam penggolongan teks, yaitu clustering teks dan klasifikasi teks. *Clustering* teks berhubungan dengan menemukan sebuah struktur kelompok yang belum kelihatan (tak terpandu atau unsupervised) dari sekumpulan dokumen. Sedangkan pengklasifikasian teks dapat dianggap sebagai proses untuk membentuk golongan-golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya (terpandu atau *supervised*).

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).



Gambar 2.1. Pembelajaran mesin dengan klasifikasi terpadu

Gambar 2.1. menunjukkan proses dari klasifikasi teks secara terpandu menggunakan pembelajaran mesin. Input mengalami *preprocessing* yaitu bisa berupa *stopword* atau *stemming.* selama proses pelatihan, ekstraksi fitur diterapkan untuk mengkonversi setiap nilai masukanke himpunan fitur. Pasangan himpunan fitur dan label kemudian diumpankan ke algoritma pembelajaran mesin untuk membangkitkan sebuah model. Selama prediksi ektraksi fitur yang sama diterapkan untuk mengkonversi masukan-masukan baru ke himpunan fitur. Himpunan fitur ini lalu diumpankan ke model, yang akan membangkitkan perkiraan (prediksi) label. Sewaktu pengujian, prediksi-prediksi label ini dicocokkan dengan label sebenarnya untuk mengevaluasi kinerja pengklasifikasi teks terpadu. Beberapa cara pada pengolahan teks antara lain:

1. *Information retrieval*: pencarian dokumen
2. Klasifikasi dokumen: membagi dokumen kedalam kelas-kelas yang telah ditentukan sebelumnya. Misalnya secara otomatis dapat menentukan apakah dokumen ini masuk ke dalam ketegori politik, ekonomi, militer dan lain sebagainya.
3. *Document Clustering*: mirip dengan klasifikasi dokumen, hanya saja kelas dokumen tidak ditentukan sebelumnya. Misalnya berita tentang lalulintas dapat menjadi satu kelas dengan berita tentang kriminal karena didalamnya banyak memuat tentang orang yang tewas, cedera, rumah sakit.
4. peringkasan teks: menghasilkan ringkasan suatu dokumen secara otomatis.
5. ekstraksi informasi: Mengekstrak informasi yang dianggap penting dari suatu dokumen. Misalnya pada dokumen lowongan, walaupun memiliki format beragam dapat diekstrak secara otomatis *job title*, tingkat pendidikan, penguasaan bahasa.

2.2.4. Teks *Preprocessing*

Cara yang digunakan dalam mempelajari suatu data teks, adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Sebelum menentukan fitur-fitur yang mewakili, diperlukan tahapan ekstraksi yang dilakukan secara umum pada dokumen yaitu tokenizing, filtering, stemming (Mooney, 2006).

Pembangunan *index* dari koleksi 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*).

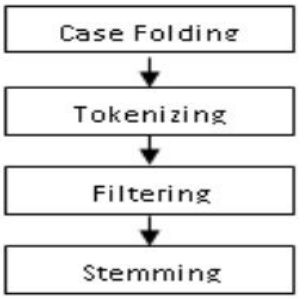
Pembuatan *inverted index* harus melibatkan konsep *linguistic processing* yang bertujuan mengesktrak *term-term* penting dari dokumen yang direpresentasikan sebagai *bag-of-words*. Ekstraksi *term* biasanya melibatkan dua operasi utama yaitu penghapusan *stop-words* dan *stemming* (Cios et al., 2007). 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).

Proses ekstraksi dokumen ditunjukkan oleh gambar 2.2. berikut penjelasan dari masing-masing komponen:

1. Tahap *case folding* adalah proses mengubah huruf kecil dalam dokumen menjadi huruf kecil. Hanya ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap sebagai delimiter.
2. Tahap *tokenizing/parsing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.
3. Tahap *filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting). *Stoplist*/*stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang.

Contoh *stopwords* adalah “yang”, “dan”, “di”, “dari” dan seterusnya.

4. Tahap *stemming* adalah tahap mencari *root* kata dari tiap kata hasil *filtering*. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki persamaan bentuk baku yang permanen.



Gambar 2.2. *Text Preprocessing*

2.2.5. Algoritma *Naïve Bayes Classifier*

Pada metode *Naïve Bayes Classifier*, sebuah dokumen teks direpresentasikan sebagai kumpulan kata-kata (*bag of words),* dimana tiap-tiap kata dalam dokumen tersebut diasumsikan tidak bergantung satu sama lain (Schneider, 2005). Dengan aturan Bayes maka dapat dinyatakan bahwa:

Dimana c_j adalah kategorisasi teks yang akan diklasifikasikan, dan $p(c_j)$ merupakan probabilitas dari kategori teks c_j. Sedangkan d merupakan dokumen teks yang dapat direpresentasikan sebagai himpunan kata (w_1, w_2 ... w_n), dimana w_1 adalah kata pertama, w_2 adalah kata kedua dan seterusnya.

Pada saat proses pengklasifikasian dokumen teks, maka pendekatan Bayes akan menyeleksi kategorisasi teks yang memiliki probabilitas paling tinggi (c_{MAP}) yaitu:

Nilai p(d) dapat diabaikan karena nilainya adalah konstan untuk semua c_j, sehingga persamaan tersebut dapat dituliskan sebagai berikut:

Probabilitas $p(c_j)$ dapat diestimasi dengan cara menghitung jumlah dokumen training pada tiap-tiap kategorisasi teks c_j. Sedangkan untuk menghitung distribusi $p(d \mid c_j)$ akan dulit untuk dilakukan khususnya pada proses pengklasifikasian dokumen teks yang berjumlah besar. Hal ini disebabkan karena jumah term tersebut sama dengan jumlah semua kombinasi posisi kata dikalikan dengan jumlah kategori yang akan diklasifikasikan.

Dengan pendekatan *Naïve Bayes Classifier* yang mengasumsikan bahwa tiap-tiap kata didalam setiap kategori adalah tidak bergantung satu sama lain, maka perhitungan dapat lebih disederhanakan dan dapat dituliskan sebagai berikut:

Dengan menggunakan persamaan diatas, maka persamaan tersebut dapat dituliskan menjadi:

Nilai $p(c_j)$ dan $p(w_i \mid c_j)$ akan dihitung pada saat proses training dijalankan yaitu:

Dimana $n(w_j)$ adalah jumlah kata pada kategori j dan n(sampel) adalah jumlah dokumen sampel yang digunakan dalam proses training. Sedangkan nj adalah jumlah kemunculan kata wi pada kategori c_j, |C| adalah jumlah semua kata pada kategori c_j dan n(kosakata) adalah jumlah kata yang unik pada semua data training.

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

Pada penelitian ini himpunan data yang akan diuji adalah kumpulan artikel-artikel yang diambil dari situs koran digital www.kompas.com serta dibagi menurut kelas-kelasnya. Kelas-kelas yang dimaksud adalah pengkategorian dari tiap jenis artikel yang disesuaikan dengan pengkategorian artikel didalam situs www.kompas.com, sehingga bisa dibedakan menjadi 3 kelas, yaitu:

1. Gosip
2. Kuliner
3. Travel

Jumlah total artikel yang digunakan pada penelitian ini adalah ±240 data teks yang tersebar pada tiap-tiap kelasnya.

Preprocessing yang dilakukkan dengan *Tokenization, Tranform Cases. 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 *Naïve Bayes Classifier*.
3. Metode Yang Diusulkan
   Untuk meningkatkan akurasi dari Algoritma *Naïve Bayes Classifier*, 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 *Naïve Bayes Classifier*. Validasi digunakan untuk melihat perbandingan hasil akurasi dari model yang digunakan dengan hasil yang telah ada sebelumnya.

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 *Naïve Bayes Classifier* yang digunakan sebagai metode pemilihan fitur agar akurasi pengklasifikasi *Naïve Bayes Classifier* bisa meningkat. Penulis menggunakan pengklasifikasi *Naïve Bayes Classifier* karena merupakan teknik *machine learning* yang populer untuk klasifikasi teks, serta memiliki performa yang baik pada banyak domain.

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.

3.6. Evaluasi dan Validasi Hasil

Model yang diusulkan pada penelitian tentang klasifikasi konten berita digital bahasa Indonesia adalah dengan menerapkan *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* (PSO). Penerapan algoritma *Naïve Bayes Classifier* 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.

4.1.1. Klasifikasi Teks Menggunakan Algoritma *Naïve Bayes Classifier*

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*
Si cantik Raisa berkesempatan untuk menghadiri acara Influence Asia 2015 yang digelar pada Senin (7/12/2015) di Suntec Convention Centre, Singapura. Dalam acara tersebut, ia bertemu banyak artis-artis dari belahan Asia lain. Salah satunya adalah Tiffany 'SNSD'.	Si cantik Raisa berkesempatan untuk menghadiri acara Influence Asia 2015 yang digelar pada Senin 7122015 di Suntec Convention Centre Singapura Dalam acara tersebut ia bertemu banyak artis artis dari belahan Asia lain. Salah satunya adalah Tiffany SNSD	si cantik raisa berkesempatan untuk menghadiri acara influence asia 2015 yang digelar pada senin 7122015 di suntec convention centre singapura dalam acara tersebut ia bertemu banyak artis artis dari belahan asia lain. salah satunya adalah tiffany snsd

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 *Naïve Bayes Classifier* 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 dokumen 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 *Naïve Bayes Classifier* 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	Naive Bayes + PSO *Accuracy*
2	90.42 %
3	91.25 %
4	92.50 %
5	94.17 %
6	92.92 %
7	91.70 %
8	92.50 %
9	94.16 %
10	92.50 %

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 *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* untuk menentukan nilai *accuracy*. Dalam menentukan nilai tingkat keakurasian dalam model *Naïve Bayes Classifier* dan algoritma *Particle Swarm Optimization.*

4.2.1. Analisis Evaluasi Hasil dan Validasi Model

Dari hasil di atas, pengukuran akurasi menggunakan *confusion matrix* terbukti bahwa hasil pengujian algoritma *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* memiliki nilai akurasi yang lebih tinggi. Dengan nilai akurasi sebesar 94.17% yang didapat saat proses 5 *fold cross validation.*

4.3. Pembahasan

Berdasarkan hasil eksperimen yang dilakukan untuk memecahkan masalah klasifikasi konten berita digital, dapat disimpulkan bahwa hasil eksperimen menggunakan metode *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* didapat nilai akurasi terbaik yaitu mempunyai akurasi sebesar 94.17%.

Tabel 4.3 Model *Confusion Matrix* Untuk Metode *Naïve Bayes Classifier* Berbasis *Particle Swarm Optimization*

Accuracy: 94.17% +/- 5.08% (mikro: 94.17%)				
	true Travel	true Gosip	true Kuliner	class precision
Pred Travel	75	1	3	94.94%
Pred Gosip	2	74	0	97.37%
Pred Kuliner	3	5	77	90.59%
Class recall	93.75 %	92.50 %	96.25%	

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 *Naïve Bayes Classifier* 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 *Naïve Bayes Classifier* berbasis *Particle 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, gunung makan, minum, restoran, lezat, kuliner. Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy, precision,* dan *recall* dari setiap algoritma sehingga didapatkan pengujian dengan menggunakan *Naïve Bayes Classifier* berbasis

Particle Swarm Optimization (PSO) didapatkan nilai *accuracy* 94.17%. Maka dapat disimpulkan pengujian data konten berita digital menggunakan *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* (PSO) sangat baik digunakan dalam klasifikasi konten berita bahasa Indonesia. Dengan demikian hasil dari pengujian model di atas dapat disimpulkan bahwa *Naïve Bayes Classifier* berbasis *Particle Swarm Optimization* (PSO) memberikan pemecahan untuk permasalahan klasifikasi konten berita digital lebih akurat.

Walaupun pengklasifikasi *Naïve Bayes Classifier* 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.

6. DAFTAR PUSTAKA

[1] 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.

[2] Chandra, D. N., Indrawan, G., & Sukajaya, I. N. (2016). Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram.

[3] Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). Data Mining A Knowledge Discovery Approach. Springer.

[4] Efendi, R., & Malik, R. F. (2012). Klasifikasi dokumen berbahasa indonesia menggunakan naive bayes classifier.

[5] Feldman, R., & Sanger, J. (2007). The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press.

[6] Gorunescu, F. (2011). *Data Mining Concept Model Technique*.

[7] Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Soft Computing* (Vol. 54).

[8] Haupt, S. E. (2004). *Practical Genetic Algorithms*.

[9] Kaizhu Huang, Haiqin Yang, Irwin King, M. L. (2008). *Advanced Topics in Science and Technology in China*.

[10] Kaur, H. (2013). Online News Classification : A Review, 7–9.

[11] 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.

[12] 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.

[13] Liu, Y., Wang, G., Chen, H., & Dong, H. (2011). An improved particle swarm optimization for feature selection. *Journal of Bionic …*, *8*, 392–397.

[14] Mahinovs, A. Tiwarigton, a. (2007). Text Classification Method Review. *Decision Engineering Report Series*, (April).

[15] Maimon, O. (2010). *Data Mining And Knowledge Discovery Handbook.* New York Dordrecht Heidelberg London: Springer.

[16] Manning, C. D., Ragnavan, P., & Schutze, H. (2008). An Introduction to Information Retreival. *IEEE Photonics Technology Letters*, *21*(8), C3–C3.

[17] Mooney, J. (2006). Machine Learning Text Categorization. Austin: University of Texas.

[18] Polettini, N. (2004). The Vector Space Model in Information Retrieval - Term Weighting Problem. 1-9.

[19] Samodra, J., Sumpeno, S., & Hariadi, M. (2009). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes.

[20] Schneider, Karl-Michael. (2005). Techniques for Improving the Performance of Naive Bayes for Text Classification. In Proceedings of CICLing, pages 682-693.

[21] Sebastiani, F. (2001). Machine Learning in Automated Text Categorization.