**PROPOSAL TUGAS AKHIR**

**PENERAPAN *NAÏVE BAYES* (NB) DAN *PARTICLE SWARM OPTIMIZATION* (PSO) PADA SISTEM KLASIFIKASI CERITA PENDEK**

****

**NI PUTU ARY RARA ISWARI**

**NIM. 1508605016**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN**

**UNIVERSITAS UDAYANA**

**BUKIT JIMBARAN**

**2018**

# **LEMBAR PENGESAHAN PROPOSAL TUGAS AKHIR**

Judul : Penerapan *Naïve Bayes* (NB) dan *Particle Swarm Optimization* (PSO) Pada Sistem Klasifikasi Cerita Pendek

Kompetensi : Komputasi

Nama : Ni Putu Ary Rara Iswari

NIM : 1508605016

Tanggal Seminar : 28 November 2018

Disetujui oleh:

|  |  |
| --- | --- |
| Reviewer I  Dr. A. A. I. N. Eka Karyawati, S.Si., M.Eng.  NIP. 19740404 199802 2 001 | Penguji I  Dr. I Ketut Gede Suhartana, S.Kom., M.Kom.  NIP. 19720110 200812 1 001 |
| Reviewer II  I Gede Santi Astawa, S.T., M.Cs.  NIP. 19821206 200604 1 003 | Penguji II  Dra. Luh Gede Astuti, M.Kom.  NIP. 19640114 199402 2 001 |
|  | Penguji III  I Dewa Made Bayu A. D., S.Kom., M.Cs.  NIP. 19890127 201212 1 001 |

Mengetahui,

Komisi Seminar dan Tugas Akhir

Jurusan Ilmu Komputer FMIPA UNUD

I Gede Arta Wibawa, S.T., M.Kom.

NIP. 198310222008121001

# **KATA PENGANTAR**

Proposal penelitian dengan judul ini disusun dalam rangkaian kegiatan pelaksanaan Tugas Akhir di Jurusan Ilmu Komputer FMIPA UNUD. Proposal ini disusun dengan harapan dapat menjadi pedoman dan arahan dalam melaksanakan penelitian di atas.

Sehubungan dengan telah terselesainya proposal ini, maka diucapkan terima kasih dan penghargaan kepada berbagai pihak yang telah membantu pengusul, antara lain:

1. Ibu Dr. A. A. I Ngr Eka Karyawati, S.Si., M.Eng. sebagai calon Pembimbing I yang telah banyak membantu menyempurnakan proposal ini;
2. Bapak I Gede Santi Astawa, S.T., M.Cs. sebagai calon Pembimbing II yang telah bersedia mengkritisi, memeriksa dan menyempurnakan proposal ini;
3. Bapak-bapak dan Ibu-ibu dosen pengajar di Jurusan Ilmu Komputer Fakultas MIPA Universitas Udayana yang telah meluangkan waktu untuk memberikan saran dan masukan dalam menyempurnakan proposal ini;
4. Kawan-kawan di Jurusan Ilmu Komputer yang telah memberikan dukungan moral dalam penyelesaian proposal ini.

Disadari pula bahwa sudah tentu proposal ini masih mengandung kelemahan dan kekurangan. Memperhatikan hal ini, maka masukan dan saran-saran penyempurnaan sangat diharapkan.

Bukit Jimbaran, Desember 2018

Penyusun

# **DAFTAR ISI**

[LEMBAR PENGESAHAN PROPOSAL TUGAS AKHIR ii](#_Toc533154766)

[KATA PENGANTAR iii](#_Toc533154767)

[DAFTAR ISI iv](#_Toc533154768)

[DAFTAR TABEL](#_Toc533154768) vi

[DAFTAR GAMBAR vii](#_Toc533154768)

[1. Latar Belakang 1](#_Toc533154769)

[2. Rumusan Masalah 3](#_Toc533154770)

[3. Tujuan Penelitian 3](#_Toc533154771)

[4. Batasan Masalah 3](#_Toc533154772)

[5. Manfaat Penelitian 4](#_Toc533154773)

[6. Tinjauan Pustaka 4](#_Toc533154774)

[6.1. Tinjauan Empiris 4](#_Toc533154775)

[6.2. Tinjauan Teoritis 5](#_Toc533154776)

[6.2.1. Klasifikasi Dokumen 5](#_Toc533154777)

[6.2.2. Cerita Pendek 6](#_Toc533154778)

[6.2.3. Text Preprocessing 6](#_Toc533154779)

[6.2.4. Pembobotan TF-IDF 9](#_Toc533154780)

[6.2.5. Seleksi Fitur PSO 10](#_Toc533154781)

[6.2.6. Naïve Bayes 10](#_Toc533154782)

[6.2.7. Evaluasi Kelas 11](#_Toc533154783)

[7. Metodologi Penelitian 12](#_Toc533154784)

[7.1. Pengumpulan Data 12](#_Toc533154785)

[7.2. Alur Penelitian 13](#_Toc533154786)

[7.3. Tahap Preprocessing 13](#_Toc533154787)

[7.3.1. Transform Cases 13](#_Toc533154788)

[7.3.2. Tokenized 14](#_Toc533154789)

[7.3.3. Filtering 14](#_Toc533154790)

[7.3.4. Stemming 15](#_Toc533154791)

[7.4. Seleksi Fitur PSO 15](#_Toc533154792)

[7.4.1. Inisialisasi Kecepatan Awal 15](#_Toc533154793)

[7.4.2. Inisialisasi Posisi Awal Partikel 15](#_Toc533154794)

[7.4.3. Inisialisasi Pbest dan Gbest Awal 16](#_Toc533154795)

[7.4.4. Update Kecepatan 16](#_Toc533154796)

[7.4.5. Update Posisi 16](#_Toc533154797)

[7.5. Naïve Bayes 17](#_Toc533154798)

[7.5.1. Learning (Pembelajaran) 17](#_Toc533154799)

[7.5.2. Classify (Pengklasifikasian) 18](#_Toc533154800)

[7.6. Pengujian 18](#_Toc533154801)

[8. Jadwal Pelaksanaan Penelitian 19](#_Toc533154802)

[DAFTAR PUSTAKA 20](#_Toc533154803)

**DAFTAR TABEL**

[Tabel 1 Contoh Potongan Teks Cerita Pendek 12](#_Toc515808412)

[Tabel 2 Contoh Transform Cases (ToLowerCase) 13](#_Toc515808413)

[Tabel 3 Contoh Tokenized 14](#_Toc515808414)

[Tabel 4 Contoh Proses Stopword 14](#_Toc515808415)

[Tabel 5 Contoh Stemming 15](#_Toc515808416)

**DAFTAR GAMBAR**

[Gambar 1 Alur Penelitian 13](#_Toc515808479)

# **Latar Belakang**

Indonesia adalah negara yang sangat besar dapat dilihat mulai dari jumlah penduduk, luas wilayah, sumber daya alam hingga seni budaya dan adat istiadatnya. Apabila dilihat lebih mendalam dari sisi budaya, bahwa budaya adalah suatu cara hidup yang berkembang dan dimiliki bersama oleh sebuah kelompok orang serta diwariskan dari generasi ke generasi. Budaya dapat terbentuk dari berbagai unsur diantaranya mulai dari agama, politik, adat istiadat, bahasa, pakaian, hingga karya seni. Salah satu hasil dari kebudayaan adalah karya sastra.

Salah satu karya sastra Indonesia yang diminati oleh banyak orang adalah cerita pendek (cerpen). Cerpen adalah cerita yang membatasi diri dalam membahas salah satu unsur fiksi dalam aspeknya yang terkecil. Kependekan sebuah cerita pendek bukan karena bentuknya yang jauh lebih pendek dari novel, tetapi karena aspek masalahnya yang sangat dibatasi (Sumardjo, 1983: 69). Untuk saat ini terdapat beberapa kategori dari cerita pendek diantaranya adalah cerpen anak, dongeng, fabel, fantasi, dan masih banyak lagi. Menurut salah seorang penulis, Eka Kurniawan bahwa, sejak tahun 1999 beliau telah menulis lebih dari 50 cerpen dan menurutnya itu pun lebih dari cukup. Melihat hal tersebut dari seorang penulis saja dapat menulis 50 cerpen dalam kurun waktu kurang lebih 18 tahun. Bila seluruh penulis di Indonesia menulis cerpen dengan jumlah rata-rata yang sama, maka akan begitu banyak cerpen yang ada di Indonesia. Namun, dari cukup banyaknya cerpen yang telah diterbitkan pada media massa biasanya masih banyak dari cerpen tersebut yang belum dapat diketahui kategorinya. Untuk bisa mengetahui kategori dari cerpen tersebut dibutuhkan waktu yang cukup lama karena pembaca harus membaca setidaknya paling sedikit setengah dari panjang cerita tersebut. Dengan demikian diperlukan suatu teknologi yang dapat mempermudah dalam mengkategorikan cerpen tersebut.

Dalam menentukan kategori dari cerpen dapat dilakukan dengan proses pengklasifikasian. Terdapat beberapa metode yang dapat digunakan dalam menyelesaikan masalah klasifikasi diantaranya *Decision Tress*, *k-Nearest Neighbors* (kNN), *Neural Networks*, *Naïve Bayes* (NB), dan *Support Vector Machines* (SVM). Salah satu metode yang cocok dalam pengklasifikasian teks atau dokumen adalah metode *Naïve Bayes* (NB) yang memiliki beberapa kelebihan antara lain, sederhana, cepat dan berakurasi tinggi. Sehingga penelitian yang akan penulis lakukan adalah membuat sistem yang mampu mengklasifikasikan kategori cerita pendek dengan menggunakan metode *Naïve Bayes* (NB).

Terdapat beberapa penelitian yang sudah dilakukan terkait sistem pengklasifikasian. Somantri (2017) melakukan 2 penelitian diantaranya klasifikasi cerita pendek berbahasa Indonesia dengan text mining menggunakan metode Naïve Bayes dan melakukan seleksi fitur dalam klasifikasi cerita pendek menggunakan Algoritma Genetika. Penelitian lainnya dilakukan oleh Erfian dan Dwiza (2017), dimana penelitian tersebut melakukan klasifikasi dokumen berita berbahasa Inggris menggunakan Naïve Bayes Classifier. Dalam penelitian tersebut digunakan PSO untuk seleksi fitur dengan sebelumnya dilakukan perhitungan TF-IDF.

Fokus dari penelitian yang ingin dilakukan penulis adalah pengembangan dari penelitian yang telah dilakukan Somantri (2017) dengan tetap menggunakan metode Naïve Bayes sebagai metode klasifikasi yang memiliki kelemahan, yaitu sangat sensitif terhadap pemilihan fitur seleksi. Oleh karena itu, sebuah pemilihan fitur yang sesuai dengan model yang diusulkan sangatlah diperlukan. Sehingga digunakan algoritma optimasi PSO sebagai seleksi fitur karena algoritma ini mampu meningkatkan *precision* dari *term* extraction (Syafrullah & Salim, 2010).

Dalam penelitian ini fitur yang digunakan berupa kata-kata dari cerpen tersebut yang sudah menjadi term dengan setiap fitur dibobotkan dengan TF-IDF. Selain itu klasifikasi kategori cerpen dibagi menjadi 3 kelas yaitu cerpen anak, fantasi dan tidak diketahui. Cerpen anak merupakan cerita tentang kehidupan anak baik suka dukanya dalam keluarga dan masyarakat (Azka, 2015). Sedangkan cerpen fantasi merupakan cerita yang dibangun dalam alur penceritaan yang normal namun bersifat imajinatif dan hayalan karena biasanya dalam setting, penokohan, maupun konflik tidak realistis bahkan terkesan dilebih-lebihkan dan tidak mungkin terjadi di dunia nyata (Kemendikbud RI, 2016). Pada kelas tidak ketahui ini merupakan klasifikasi kategori cerpen diluar kelas cerpen anak dan fantasi seperti cerpen dongeng, fabel dan yang lainnya. Penelitian yang akan penulis lakukan berguna dalam hal otomatisasi klasifikasi agar pengarsipan data berupa cerita pendek berbahasa Indonesia seperti pada perpustakaan elektronik khususnya yang mengarsipkan karya sastra berupa cerita pendek ini dapat dilakukan dengan lebih efektif dan efisien. Hal tersebut menjadi penting karena di zaman modern ini banyak pekerjaan sudah dilakukan dengan terkomputerisasi.

# **Rumusan Masalah**

Dari latar belakang penelitian, dapat dirumuskan permasalahan pada penelitian ini adalah sebagai berikut.

1. Bagaimana Tingkat Akurasi Metode Naïve Bayes dan PSO Untuk Klasifikasi Cerita Pendek ?
2. Berapa Peningkatan Akurasi Terhadap Metode Naïve Bayes dan Metode Naïve Bayes – PSO ?

# **Tujuan Penelitian**

Adapun tujuan dari penelitian ini adalah sebagai berikut.

1. Mengetahui Tingkat Akurasi Metode Naïve Bayes dan PSO Untuk Klasifikasi Cerita Pendek
2. Mengetahui Peningkatan Akurasi Terhadap Metode Naïve Bayes dan Metode Naïve Bayes – PSO

# **Batasan Masalah**

Adapun batasan masalah dari penelitian ini adalah sebagai berikut.

1. Cerita pendek yang digunakan berbahasa Indonesia
2. Dataset (cerita pendek) yang digunakan memiliki format .txt
3. Sistem hanya dapat mengklasifikasikan cerita pendek ke dalam 3 class, yaitu cerita pendek anak, fantasi, dan tidak diketahui
4. Sistem akan dikembangkan berbasis web agar dapat diakses melakukan perangkat mobile ataupun komputer

# **Manfaat Penelitian**

Adapun manfaat yang akan dicapai pada penelitian ini adalah sebagai berikut.

1. Menambah wawasan mengenai proses optimasi dalam klasifikasi menggunakan Naïve Bayes
2. Pemanfaatan teknologi computer dalam klasifikasi kategori cerita pendek berbahasa Indonesia dengan waktu yang lebih singkat

# **Tinjauan Pustaka**

## **Tinjauan Empiris**

Pada penelitian ini, terdapat beberapa penelitian terkait mengenai kasus klasifikasi teks antara lain.

1. ***Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB).***

(Somantri. 2017)

Pada jurnal ini dibahas metode NB digunakan untuk klasifikasi cerita pendek. Metode NB sebagai metode yang diusulkan dan diterapkan karena merupakan salah satu metode komputasi yang efisien dan mempunyai *performance predictive* yang baik sehingga menjadi metode klasifikasi teks yang popular. Metode ini di evaluasi dengan membandingkan tingkat akurasi dengan metode *Support Vector Machine* (SVM). Hasil akurasi yang diperoleh sebesar 78,59% lebih besar dibandingkan dengan metode SVM.

1. ***Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritma Genetika.***

(Somantri. 2017)

Jurnal ini merupakan pengembangan jurnal sebelumnya yaitu pengklasifikasian kategori cerpen dengan menggunakan *text mining*. Fokus penelitian ini adalah peningkatan akurasi klasifikasi teks pada NB dengan menggunakan pendekatan *feature selection*. Algoritma genetika adalah salah satu algoritma optimasi yang dapat digunakan untuk permasalahan *feature selection*. Evaluasi dalam penelitian ini dilakukan dengan menggunakan komparasi antara tingkat akurasi model NB klasik, SVM, NB-GA dan SVM-PSO. Akurasi yang dihasilkan sebesar 84,29%.

1. ***Penerapan PSO Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC.***

(Erfian, Dwiza. 2017)

Pada penelitian ini dilakukan penerapan algoritma PSO untuk seleksi fitur pada klasifikasi dokumen karena pada penelitian sebelumnya belum ada yang menerapkan. Algoritma ini mampu mengurangi dimensi dokumen yang sangat besar sekaligus mampu meningkatkan akurasi. Metode klasifikasi yang digunakan adalah Naïve Bayes Classifier (NBC). Algoritma ini juga mampu menangani klasifikasi dengan jumlah data yang besar dan kategori yang cukup banyak. Hasil akurasi dari data training adalah 85,42% dan dari data testing adalah 99,87%

## **Tinjauan Teoritis**

### **Klasifikasi Dokumen**

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Dalam mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu membedakan data ke dalam kelas-kelas yang berbeda berdasarkan aturan-aturan atau fungsi tertentu. Model itu sendiri dapat berupa aturan “jika-maka”, berupa pohon keputusan, ataupun formula matematis (Bustami, 2014).

Klasifikasi dokumen adalah pemberian kategori yang telah didefinisikan kepada dokumen yang belum memiliki kategori (Goller, 2000). Mengklasifikasi dokumen merupakan salah satu cara untuk mengorganisasikan dokumen. Dokumen-dokumen yang memiliki isi yang sama akan dikelompokkan ke dalam kategori yang sama. Dengan demikian, orang-orang yang melakukan pencarian informasi dapat dengan mudah melewatkan kategori yang tidak relevan dengan informasi yang dicari atau yang tidak menarik perhatian (Feldman, 2004).

### **Cerita Pendek**

Pengertian cerita pendek kini telah banyak dikemukakan pakar sastra dan sastrawan. Meskipun sulit untuk dijelaskan lebih rinci mengenai definisi dari cerita pendek, namun berikut itu terdapat beberapa pengertian cerita pendek menurut para pakar sastra maupun sastrawan terkemuka.

Cerita pendek adalah cerita yang membatasi diri dalam membahas salah satu unsur fiksi dalam aspeknya yang terkecil. Kependekan sebuah cerita pendek bukan karena bentuknya yang jauh lebih pendek dari novel, tetapi karena aspek masalahnya yang sangat dibatasi (Sumardjo, 1983: 69).

Sesuai dengan namanya, cerita pendek dapat diartikan sebagai cerita berbentuk prosa yang pendek (Suyanto, 2012:46). Ukuran pendek di sisni bersifat relatif. Menurut Edgar Allan Poe dalam (Suyanto, 2012:46), sastrawan kenamaan Amerika, ukuran pendek di sini adalah selesai dibaca dalam sekali duduk, yakni kira-kira kurang dari satu jam. Adapun Jacob Sumardjo dan Saini K.M (1995: 30) dalam Suyanto (2012: 46) menilai ukuran pendek ini lebih didasarkan pada keterbatasan pengembangan unsur-unsurnya. Cerpen harus memiliki efek tunggal dan tidak kompleks.

Berdasarkan beberapa pengertian tersebut Rindi (2017) dapat mengambil menyimpulkan bahwa pengerti cerita pendek adalah sebuah karangan berbentuk prosa fiksi yang habis dbaca sekali duduk, maksud dari habis dibaca sekali duduk adalah tidak membutuhkan waktu yang berlama-lama untuk menyelesaikan satu cerita. Cerita pendek juga memiliki pemendekan unsur-unsur pembentuknya, jadi kaya akan pemadatan makna.

### **Text Preprocessing**

Tahap *text pre-processing* adalah tahap awal dari *text mining.* Tahap ini mencakup semua rutinitas, dan proses untuk mempersiapkan data yang akan digunakan pada operasi *knowledge discovery* sistem *text mining* (Fieldman & Sanger, 2007). Setelah dokumen diinputkan dalam sistem maka akan dilakukan beberapa langkah-langkah *text pre-processing* yang akan dijabarkan seperti berikut ini (Tala, 2003).

1. Transform Cases

Tahap ini merupakan proses untuk mengubah bentuk kata dalam dokumen. Pada proses ini, semua huruf dalam dokumen dijadikan huruf kecil (to *lower case*) dalam hal ini hanya huruf ‘a’ sampai dengan ‘z’ yang dapat diterima. Berikut contoh dari proses transform cases.

|  |  |
| --- | --- |
| Input | Output |
| KOMPUTER | komputer |
| KomPUter | komputer |
| manaJEMEN | manajemen |
| MaNaJemen | manajemen |

1. Tokenized

Tahap ini merupakan proses pemotongan string input berdasarkan tiap kata penyusunnya. Pada prinsipnya proses ini adalah memisahkan setiap kata yang menyusun suatu dokumen. Secara umum karakter whitespace, seperti enter, tabulasi, spasi dianggap sebagai pemisah kata. Pada proses ini dilakukan pula penghilangan angka, tanda baca dan karakter selain huruf alfabet, karena karakter-karakter tersebut dianggap sebagai pemisah kata (delimiter) dan tidak memiliki pengaruh terhadap pemrosesan teks. Berikut contoh dari proses tokenized.

|  |  |
| --- | --- |
| Teks Input | Hasil Proses Tokenized |
| manajemen pengetahuan adalah sebuah konsep baru di dunia bisnis. | manajemen  pengetahuan  adalah  sebuah  konsep  baru  di  dunia  bisnis |

1. Filtering

Tahap f**iltering** adalah tahap mengambil kata-kata penting dari hasil token. Tahap ini bisa menggunakan **algoritma stoplist (membuang kata kurang penting)** atau **wordlist (menyimpan kata penting)**.  Stoplist / stopword merupakan kata-kata yang tidak deskriptif (tidak penting) yang dapat dibuang dengan pendekatan bag-of-words. Pada tahapan ini setelah diperoleh token dari seluruh dokumen kemudian apabila hasil tokenisasi tersebut ada yang merupakan kata tidak penting dalam database maka hasil tokenisasi tersebut dibuang. Berikut contoh dari proses filtering.

|  |  |
| --- | --- |
| Hasil Tokenisasi | Hasil Filtering |
| manajemen  pengetahuan  adalah  sebuah  konsep  baru  di  dunia  bisnis | manajemen  pengetahuan  konsep  baru  dunia  bisnis |

1. Stemming

Tahap ini merupakan proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari root kata dari tiap kata hasil filtering. Dengan dilakukannya proses stemming setiap kata berimbuhan akan berubah menjadi kata dasar, dengan demikian dapat lebih mengoptimalkan hasil pada proses selanjutnya. Pengimplementasian proses stemming sangatlah beragam yang bergantung dengan bahasa dari dokumen. Berikut contoh dari proses stemming.

|  |  |
| --- | --- |
| Hasil Filtering | Hasil Stemming |
| membela  menguatnya  dikatakan  dibandingkan | bela  kuat  kata  banding |

### **Pembobotan TF-IDF**

Pembobotan dapat diperoleh berdasarkan jumlah kemunculan suatu *term* dalam sebuah dokumen *term frequency* (*tf*) dan jumlah kemunculan *term* dalam koleksi dokumen *inverse document frequency* (*idf*). Bobot suatu istilah semakin besar jika istilah tersebut sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen (Grossman, 1998). Nilai *idf* sebuah *term* (kata) dapat dihitung menggunakan persamaan sebagai berikut:

………………………………………………...(1)

D adalah jumlah dokumen yang berisi *term* (*t*) dan *dfi* adalah jumlah kemunculan (frekuensi) *term* terhadap D. Adapun algoritma yang digunakan untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci (*query*), yaitu:

…………………………………………..…(2)

Keterangan :

d = dokumen ke–d

t = *term* ke–t dari kata kunci

tf = *term* frekuensi/frekuensi kata

W = bobot dokumen ke–d terhadap term ke–t

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan (*sorting*) dimana semakin besar nilai W, semakin besar tingkat kesamaan (*similarity)* dokumen tersebut terhadap kata yang dicari, demikian pula sebaliknya.

### **Seleksi Fitur PSO**

*Particle swarm optimization* adalah suatu algoritma yang banyak terinspirasi dari perilaku sosial hewan seperti burung, lebah dan ikan. Seekor hewan dalam algoritma PSO akan dianggap sebagai partikel. Partikel ini akan dipengaruhi oleh kecerdasan dari individu hewan itu sendiri dan dan kecerdasan dari partikel lain dalam satu kelompok. Apabila satu partikel menemukan jalan yang tepat dan terpendek menuju ke suatu sumber makanan, maka yang terjadi adalah partikel-partikel lain tersebut akan mengikuti partikel yang telah menemukan jalan yang tepat dan terpendek tadi.

Terdapat beberapa kondisi berhenti dalam PSO merupakan syarat yang digunakan dalam mengakhiri iterasi pencarian. Syarat-syarat tersebut akan dijelaskan sebagai berikut (Hakim, 2017) :

1. Suatu iterasi akan terhenti jika telah mencapai nilai maksimum.
2. Suatu iterasi akan terhenti jika telah ditemukan solusi yang telah memenuhi kriteria yang ada.
3. Suatu iterasi akan terhenti apabila tidak adanya perubahan nilai (konvergen).
4. Suatu iterasi akan terhenti apabila nilai radius *swarm* yang telah dinormalisasi mendekati 0.
5. Suatu iterasi akan terhenti apabila grafik fungsi obyektif telah mendekati nilai 0 seiring dengan pertambahan iterasi.

### **Naïve Bayes**

*Naïve Bayes Classifier* (NBC) merupakan salah satu metode *machine learning* yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. *Naïve Bayesian Classification* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network. Naïve Bayesian Classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar.

Teorema Bayes memiliki bentuk umum sebagai berikut (Jananto, 2013) :

…………………………………………(3)

Dalam hal ini :

*X* = Data dengan *class* yang belum diketahui

*H* = Hipotesis data X merupakan suatu *class* spesifik

*P(H|X)* = Probabilitas hipotesis H berdasar kondisi X *(posteriori*

*probability)*

*P(H) =* Probabilitas hipotesis H *(prior probability)*

*P(X|H)* = Probabilitas X berdasar kondisi pada hipotesis H

*P(X)* = Probabilitas dari X

### **Evaluasi Kelas**

Terdapat beberapa metode evaluasi yang dapat dilakukan dalam klasifikasi diantaranya adalah *precision, recall*, dan akurasi. *Precision* adalah **tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem**. Sedangkan recall adalah **tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Pada akurasi ini diartikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual (putubuku, 2008). Dengan rumus seperti berikut ini.**

………………..(4)

……………………(5)

…………………………………….(6)

# **Metodologi Penelitian**

Pada bagian metodelogi penelitian ini menjelaskan gambaran langkah-langkah yang akan dilakukan dalam menjalankan penelitian ini, langkah-langkah tersebut meliputi pengumpulan data, alur metodologi penelitian, tahap preprocessing, tahap seleksi fitur, tahap klasifikasi cerpen dan tahap pengujian.

## **Pengumpulan Data**

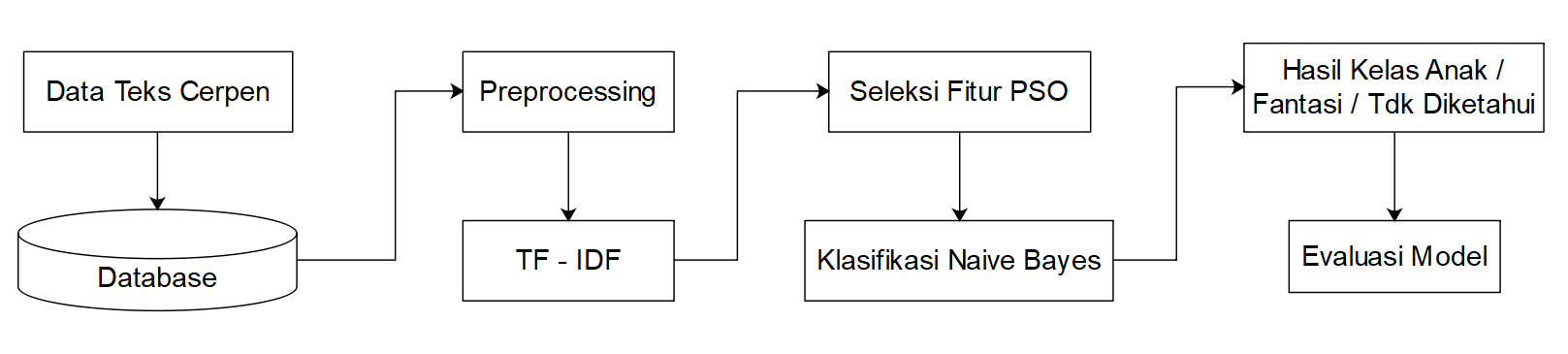
Pada pengumpulan data, data yang digunakan berupa teks cerita pendek berbahasa Indonesia yang diperoleh dari website cerpen online yaitu www.cerpenmu.com. Data untuk setiap teks cerita pendek akan disimpan dalam file dokumen cerpen dengan tipe file berekstensi .txt. Dimana pada data tersebut akan terbagi menjadi data training dan data testing yang akan diberikan label apakah termasuk kategori cerpen anak, fantasi atau tidak diketahui berdasarkan kategori cerita yang terdapat pada website. Data yang akan digunakan sebanyak 500 data dimana sebanyak 350 data digunakan sebagai data training dan 150 data digunakan sebagai data testing. Setelah data-data tersebut sudah terkumpul maka akan dimasukkan ke dalam database. Berikut merupakan contoh data cerpen anak, fantasi dan tidak diketahui.

**Tabel 1 Contoh Potongan Teks Cerita Pendek**

|  |  |  |
| --- | --- | --- |
| No. | Potongan Teks Cerpen | Kategori |
| 1 | Aku bercita cita menjadi seorang dokter. Saat ada orang yang sakit aku akan mengobatinya sampai sembuh. Kalau ada orang yang tidak punya uang untuk berobat tetap akan kuobati. Karena kesehatan itu sangatlah mahal. | Anak |
| 2 | Seorang anak berjalan menyusuri jalan yang sunyi menuju stasiun kereta untuk kembali pulang ke desanya. Di perjalanan, sesekali ia bersiul keras dengan nada yang menusuk hati siapapun yang mendengarnya dan memecahkan kesunyian di kota kecil itu. | Fantasi |
| 3 | Tersebutlah sebuah kerajaan makmur dan sejahtera, yang dipimpin oleh seorang Raja arif nan bijaksana. Sang Raja mempunyai empat orang pangeran. Dengan membawa sifat dan keahlian mereka masing-masing, keempatnya lahir ke dunia pada hari yang sama. | Tidak Diketahui |

## **Alur Penelitian**

Pada bagian ini akan digambarkan alur secara umum dari penelitian yang akan dilakukan penulis, yaitu dimulai dari pengumpulan data teks cerpen yang kemudian diekstrak ke dalam database, lalu data tersebut akan dilakukan tahap preprocessing yang kemudian dilakukan penghitungan bobot dengan TF-IDF, selanjutnya dilakukan seleksi fitur untuk mengurangi fitur yang tidak penting, setelah itu dilakukan proses klasifikasi dan akan menghasilkan kelas cerpen anak, fantasi atau tidak diketahui, dan pada tahap akhir akan dilakukan evaluasi terhadap kinerja metode yang digunakan. Penjelasan lebih detail untuk setiap langkah-langkah ini akan dijelaskan pada sub selanjutnya. Adapun alur metodologi penelitian ini sebagai berikut.



**Gambar 1 Alur Penelitian**

## **Tahap Preprocessing**

Data teks cerpen yang sudah dikumpulkan akan di olah terlebih dahulu sebelum masuk dalam tahap seleksi fitur dan tahap klasifikasi. Tahap preprocessing yang dilakukan meliputi transform cases (ToLowerCase), tokenized, stopword dan stemming.

### **Transform Cases**

Transform case merupakan proses mengkonversikan keseluruhan teks dalam dokumen menjadi bentuk huruf kecil. Berikut ini merupakan contoh hasil dari penerapan tahap transform cases.

**Tabel 2 Contoh Transform Cases (ToLowerCase)**

|  |  |
| --- | --- |
| Potongan Teks Cerpen | Hasil |
| Aku bercita cita menjadi seorang dokter. Saat ada orang yang sakit aku akan mengobatinya sampai sembuh. Kalau ada orang yang tidak punya uang untuk berobat tetap akan kuobati. Karena kesehatan itu sangatlah mahal. | aku bercita cita menjadi seorang dokter. saat ada orang yang sakit aku akan mengobatinya sampai sembuh. kalau ada orang yang tidak punya uang untuk berobat tetap akan kuobati. karena kesehatan itu sangatlah mahal. |

### **Tokenized**

Tokenized merupakan proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Berikut ini merupakan contoh hasil dari penerapan tahap tokenized.

**Tabel 3 Contoh Tokenized**

|  |  |  |  |
| --- | --- | --- | --- |
| Potongan Teks Cerpen | Hasil | | |
| Aku bercita cita menjadi seorang dokter. Saat ada orang yang sakit aku akan mengobatinya sampai sembuh. Kalau ada orang yang tidak punya uang untuk berobat tetap akan kuobati. Karena kesehatan itu sangatlah mahal. | aku  bercita  cita  menjadi  seorang  dokter  saat  ada  orang  yang  sakit  aku | akan  mengobatinya  sampai  sembuh  kalau  ada  orang  yang  tidak  punya  uang | untuk  berobat  tetap  akan  kuobati  karena  kesehatan  itu  sangatlah  mahal |

### **Filtering**

Filtering merupakan proses mengambil kata-kata penting dari hasil token. Berikut ini merupakan contoh hasil dari penerapan tahap filtering.

**Tabel 4 Contoh Proses Filtering**

|  |  |  |
| --- | --- | --- |
| Potongan Teks Cerpen | Hasil | |
| Aku bercita cita menjadi seorang dokter. Saat ada orang yang sakit aku akan mengobatinya sampai sembuh. Kalau ada orang yang tidak punya uang untuk berobat tetap akan kuobati. Karena kesehatan itu sangatlah mahal. | bercita  cita  dokter  orang  sakit  mengobatinya  sembuh | orang  uang  berobat  kuobati  kesehatan  mahal |

### **Stemming**

Stemming merupakan proses pemetaan atau penguraian bentuk dari suatu kata menjadi bentuk kata dasarnya. Berikut ini merupakan contoh hasil dari penerapan tahap stemming.

**Tabel 5 Contoh Stemming**

|  |  |  |
| --- | --- | --- |
| Potongan Teks Cerpen | Hasil | |
| Aku bercita cita menjadi seorang dokter. Saat ada orang yang sakit aku akan mengobatinya sampai sembuh. Kalau ada orang yang tidak punya uang untuk berobat tetap akan kuobati. Karena kesehatan itu sangatlah mahal. | cita  cita  dokter  orang  sakit  obat  sembuh | orang  uang  obat  obat  sehat  mahal |

## **Seleksi Fitur PSO**

Seleksi fitur dengan PSO digunakan pada data training dan data testing untuk mencari nilai optimum berdasarkan nilai fitness yang dihasilkan oleh partikel sebagai representasi solusi seleksi fitur. Nilai hasil seleksi fitur tersebut akan masuk pada table pengetahuan yang akan digunakan sebagai referensi pada proses klasifikasi dengan Naïve Bayes. Adapun proses perhitungan yang harus dilalui seperti berikut ini.

### **Inisialisasi Kecepatan Awal**

Pada saat inisialisasi kecepatan iterasi awal ini (iterasi ke-0) semua kecepatan partikel haruslah bernilai = 0 atau bisa dilihat seperti di bawah ini.

…………………………………………………………….(7)

### **Inisialisasi Posisi Awal Partikel**

Posisi awal dari sebuah partikel seharusnya berada pada range . Jika posisi yang dihasilkan kurang dari maka posisi harus dikembalikan lagi ke , dan jika posisi melebihi dari maka posisi harus dikembalikan lagi ke . Formula untuk melakukan inisialisasi sebuah partikel dengan range dimensi seperti di bawah ini.

…………………………………………(8)

### **Inisialisasi Pbest dan Gbest Awal**

Pada inisialisasi *Pbest* iterasi ke-0, nilai partikel ini dapat disamakan nilainya dengan nilai Inisialisasi awal partikel seperti di bawah ini.

………………………………………………….(9)

### **Update Kecepatan**

*Update* kecepatan berguna dalam penentuan ke arah mana perpindahan posisi partikel yang berada pada populasi tertentu. Dimana nilai bobot inersia (*w)* telah diketahui, untuk koefisien akselerasi 1 dan 2 telah diketahui, sedangkan nilai random pada dan adalah bilangan yang dipilih secara acak dalam interval 0 sampai dengan 1.

……(10)

### **Update Posisi**

Apabila kecepatan baru telah didapatkan maka dilakukan proses *update* posisi dengan cara menghitung nilai *sigmoid* berdasarkan rumus di bawah ini.

…………………………………………………(11)

Setelah nilai sigmoid pada seluruh partikel ke-I dan dimensi ke-j didapatkan maka dilakukan perbandingan antara nilai pada update kecepatan terhadap nilai acak yang telah dibuat sebanyak partikel ke-I dan dimensi ke-j. Apabila nilai *random* lebih kecil dari nilai *sigmoid* maka posisi terbaru bernilai = 1 dan sebaliknya apabila nilai *random* lebih besar dari nilai *sigmoid* maka nilai posisi terbaru bernilai = 0. Hal tersebut dapat dilihat seperti berikut ini.

………………………(12)

## **Naïve Bayes**

Metode Naïve Bayes digunakan pada data testing untuk menguji data yang masuk memiliki kelas cerpen anak, fantasi atau tidak diketahui. Untuk memilih fitur yang digunakan akan dilihat pada table pengetahuan, jika ada kata yang tidak ada ada table pengetahuan maka akan dilakukan seleksi fitur untuk kata tersebut. Berikut 2 tahapan dari metode Naïve Bayes yang harus dilalui sebagai berikut.

### **Learning (Pembelajaran)**

Berikut ini langkah-langkah pada tahapan *learning* atau tahapan pada *data training*.

1. Membentuk *vocabulary* pada setiap dokumen *data training* dilakukan dengan proses teks preprocessing
2. Menghitung probabilitas pada setiap kategori *P(X)*

…………………………………………….(13)

Keterangan :

*P(X)* : probabilitas suatu kelas pada data training

: jumlah dokumen untuk setiap kelas

: total data training yang ada

1. Menentukan probabilitas setiap kata pada setiap kategori *P(H|X)*

…………………………………………….(14)

Keterangan :

*P(H|X)* : probabilitas kemunculan setiap kata pada suatu kelas tertentu

: jumlah sebuah kata pada suatu kelas tertentu

: jumlah dokumen untuk suatu kelas

### **Classify (Pengklasifikasian)**

Berikut ini langkah-langkah pada tahapan *classify* atau tahapan pada *data testing*.

1. Menghitung *P(X)* Π *P(H|X)* untuk setiap kategori

……(15)

Keterangan :

: probabilitas dari suatu kelas untuk sebuah data testing

: probabilitas kemunculan suatu kata pada suatu kelas

1. Menentukan kategori dengan nilai *P(X)* Π *P(H|X)* maksimal yang telah dilakukan perhitungan sebelumnya

## **Pengujian**

Pengujian sistem berfungsi untuk mengetahui kinerja dari sistem itu sendiri dalam melakukan tugasnya yaitu mengklasifikasikan. Pada penelitian ini digunakan sebanyak 150 data dokumen teks cerpen yang diperoleh secara online untuk data testing yang diberikan label sesuai dengan yang sudah tertera pada web apakah cerpen tersebut terkategori cerpen anak, fantasi atau tidak diketahui. Pada penelitian ini akan dilakukan pengujian terhadap hasil klasifikasi dari data testing dengan menghitung nilai *precision, recall* dan evaluasi. Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang mendapat klasifikasi benar dengan total data testing dengan rumus sebagai berikut. Berikut ini adalah contoh pengujian dengan menggunakan penerapan *confusion matrix* untuk 3 kelas pada penelitian ini.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Class A | Class B | Class C | Total Predict for |
| Predict A | 10 | 20 | 10 | 40 |
| Predict B | 20 | 20 | 20 | 60 |
| Predict C | 25 | 15 | 10 | 50 |
| Total Class for | 55 | 55 | 40 |  |

1. *Precision*

Perhitungan *precision* dilakukan untuk setiap kelas, berikut ini akan dicontohkan untuk Class A.

0.25

1. *Recall*

Perhitungan *recall* juga dilakukan untuk setiap kelas, berikut ini akan dicontohkan untuk Class B.

0.36

1. Akurasi

Pada perhitungan akurasi akan dihitung berdasarkan rumus persamaan 6 yang akan diterapkan seperti contoh berikut ini.

0.267

# **Jadwal Pelaksanaan Penelitian**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Kegiatan** | **Minggu Ke-** | | | | | | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** |
| Pengumpulan Data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Analisis Data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Desain dan Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Analisis Hasil Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Penulisan Laporan Akhir |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# **DAFTAR PUSTAKA**

Bustami. 2014. “Penerapan Algoritma *Naïve Bayes* Untuk Mengklasifikasikan Data Nasabah Asuransi”. Jurnal Informatika Vol. 8, No. 1, Januari 2014

Ernawati, S. 2016. “Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Analisis Sentimen Review Perusahaan Penjualan Online Menggunakan Naïve Bayes”. Jurnal Evolusi Volume 4 Nomor 1 - 2016 - lppm3.bsi.ac.id/jurnal.

Fieldman, R., & Sanger, J. (2009). *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

Ginting, S. and Trinanda, R. 2014. “Teknik Data Mining Menggunakan Metode Bayes Classifier Untuk Optimalisasi Pencarian Pada Aplikasi Perpustakaan”. Jurnal Teknologi dan Informasi UNIKOM Volume 1 No 6

Grossman, D., dan Ophir, F. 1998. *Information Retrieval: Algorithm and Heuristics.* Kluwer Academic Publisher.

Hakim, S.H.F., dkk. 2017. “Seleksi Fitur Dengan *Particle Swarm Optimization* Untuk Pengenalan Pola Wajah Menggunakan *Naive Bayes* (Studi Kasus Pada Mahasiswa Universitas Brawijaya Fakultas Ilmu Komputer Gedung A)”. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 1, No. 10, Oktober 2017, hlm. 1045-1057

Junianto, E. 2017. “Penerapan *PSO* Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan *NBC*”. Jurnal Informatika, Vol.4 No.1 April 2017, pp. 38~45 E-ISSN: 2528-2247.

Mustaqhfiri, M., dkk. 2012. “Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode *Maximum Marginal Relevance*”. [Online]. Available : https://www.researchgate.net/publication/300087962

Putubuku, “Recall & Precision,” Ilmu Perpustakaan & Informasi – diskusi dan ulasan ringkas, 27-Mar-2008. [Online]. Available: <http://iperpin.wordpress.com/2008/03/27/recall-precision/>.

Somantri, O. 2016. “Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB)”. Jurnal Telematika, vol. 12 no. 1, Institut Teknologi Harapan Bangsa, Bandung p-ISSN: 1858-2516 e-ISSN: 2579-3772

Somantri, O. 2017. “Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika”. JNTETI, Vol. 6, No. 3, Agustus 2017

Sumardjo, Jakob. 1984. *Memahami Kesusatraan*. Bandung : Penerbit Alumni.

Tala, F.Z. 2003. *A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia.* Institute for Logic, Language and Computation Universite itvan Amsterdam The Netherlands.