BAB I PENDAHULUAN

Bab ini berisikan tentang alasan peneliti mengambil permasalahan ini. Pada bab ini poin-poin yang akan dipaparkan antara lain Latar Belakang, Perumusan Masalah, Batasan Masalah, Tujuan dan Metodologi Penyelesaian Masalah yang dilakukan oleh peneliti. Diharapkan dengan adanya bab ini, pembaca akan memahami poin-poin tersebut.

1.1 Latar Belakang

Perkembangan teknologi informasi terutama internet sekarang ini semakin lama semakin cepat. Sekarang, orang dapat menikmati layanan internet dengan mudah menggunakan komputer, *smartphone*, mesin permainan, ataupun TV digital. Dalam satu hari, sekitar 40% populasi di seluruh belahan dunia mengakses internet, setidaknya dari tahun 1995 bertambah 1% setiap harinya sedangkan pada tahun 1999 sampai dengan 2013 mengalami peningkatan sepuluh kali lipat dari tahun sebelumnya. Pengguna internet mencapai satu milyar pengguna pada tahun 2005, dua milyar pengguna pada tahun 2010 dan pada tahun 2015 sudah mencapai tiga milyar lebih pengguna internet aktif. Indonesia sendiri menduduki posisi ke tiga belas dalam jumlah terbanyak dalam penggunaan internet yaitu sekitar empat puluh dua juta [4].

Film merupakan suatu media komunikasi massa yang digunakan sebagai sarana hiburan bagi masyarakat. Film cukup efektif dalam menyampaikan suatu informasi. Setiap tahun, film-film baru selalu dirilis dan sangat banyak penikmat film yang menontonnya. Hasil survei yang dilakukan oleh *Classification and Rating Administration (CARA)* tahun 2014 menunjukkan bahwa jumlah film yang dirilis di Amerika Serikat dan Kanada sekitar 707 film, naik sekitar 7 persen dari tahun 2013 yaitu sekitar 659 [5]. Hasil survei tersebut membuktikan bahwa setiap tahunnya film di dunia selalu meningkat. Dengan kemajuan internet, film yang akan dirilis maupun yang sedang tayang di bioskop sangat mudah untuk diketahui. Banyak *website* yang menyediakan tentang informasi film salah satunya imdb.com. Informasi yang disajikanpun sudah cukup lengkap dengan adanya sinopsis, *genre*, artis, produser, dan lain-lain. Akan tetapi, untuk mengkategorikan *genre* pada film masih memerlukan kemampuan manusia yaitu dengan menonton film ataupun sinopsisnya terlebih dahulu lalu mengkategorikannya. Hal tersebut membutuhkan waktu dan kemampuan kognitif manusia dalam proses pengkategoriannya.

Berdasarkan permasalahan tersebut, salah satu solusi yang dapat dilakukan adalah menggunakan *machine learning*. *Machine learning* merupakan salah satu disiplin ilmu yang biasanya dilakukan untuk pemprosesan komputasi secara otomatis. Salah satu metode yang dapat dimanfaatkan dari machine learning adalah klasifikasi [6]. Klasifikasi sendiri adalah metode yang mempelajari pola-pola dari *training data* untuk memprediksikan objek yang baru ke masing-masing kelasnya [6] yang salah satu tekniknya adalah *Naive Bayesian Multi-Label Classifier*.

Naive Bayesian Multi-Label Classifier merupakan salah satu teknik dalam metode klasifikasi yang menggunakan peluang dalam proses pengklasifikasiannya serta da-

pat mengklasifikasikan lebih dari satu kelas terhadap objek baru maka dari itu teknik ini sangat cocok untuk kasus pada penelitian ini. *Peluang* yang dimodelkan oleh teknik ini dibagi atas dua macam yaitu *prior* dan *likelihood*. Teknik ini menggganggap bahwa fitur-fitur yang ada pada data dianggap independen sehingga membuat setiap kelas memiliki fitur-fitur yang merepresentasikan mereka. Selain itu juga, teknik ini memiliki banyak keuntungan antara lain prosesnya cepat, tidak membutuhkan penyimpanan yang besar, sangat baik dalam ruang lingkup yang fitur pentingnya hampir sama, kokoh terhadap fitur yang tidak relevan serta menjadi sebuah *baseline* yang dapat dipercaya pada kasus klasifikasi terhadap data teks [7]. Teknik ini diharapkan dapat memberikan sebuah solusi untuk menyelesaikan permasalahan pengkategorian *genre* film.

1.2 Perumusan Masalah

Rumusan masalah dari tugas akhir ini adalah sebagai berikut :

1. Bagaimana cara mengkategorikan satu atau lebih *genre* pada sebuah film secara otomatis?
2. Bagaimana cara membuat *classifier* yang mampu mengkategorikan *genre* pada film secara otomatis?
3. Bagaimana cara menentukan fitur - fitur yang mencirikan sebuah *genre* film?

1.3 Batasan Masalah

Batasan masalah dari tugas akhir ini adalah sebagai berikut :

1. Sinopsis yang dikategorikan merupakan bentuk teks.
2. *Genre* yang akan diklasifikasikan pada penelitian ini mencakup seluruh *main genre* pada film [8].
3. Sinopsis menggunakan Bahasa Inggris.
4. Sinopsis memiliki setidaknya 1 kalimat.
5. Bahasa pemograman yang digunakan adalah bahasa Java.
6. Database yang digunakan adalah MySQL.

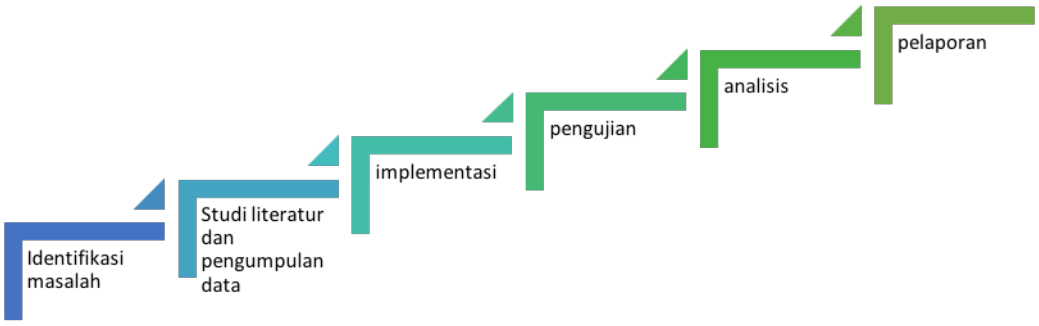
1.4 Tujuan

Tujuan dari tugas akhir ini adalah sebagai berikut :

1. Dapat mengkategorikan satu atau lebih *genre* pada film secara otomatis.
2. Dapat membangun *classifier* yang mampu mengkategorikan *genre* pada film secara otomatis.
3. Dapat menemukan fitur - fitur yang mencirikan sebuah *genre*.

1.5 Metodologi Penyelesaian Masalah

Metodologi secara umum dari tugas akhir ini dapat digambarkan pada Gambar 1.1.



Gambar 1.1: Metodologi Penyelesaian Masalah

Penjelasan untuk setiap tahapan pada Gambar 1.1 adalah sebagai berikut.

1. **Identifikasi masalah**
   Pada tahap ini peneliti mengunjungi beberapa *website* yang ada, selanjutnya peneliti mencari beberapa hal permasalahan yang bisa diberikan solusi.

2. **Studi Literatur dan pengumpulan data**
   Pada tahap ini peneliti mencari beberapa referensi jurnal/*paper* yang membahas tentang permasalahan mengenai film. Untuk pengumpulan data, peneliti melakukan pengumpulan data dengan cara mencari data di internet lalu menggunakan *interfaces* yang disediakan oleh IMDB untuk mengambil konten yang ada.

3. **Implementasi**
   Pada tahap ini ada kegiatan yang dilakukan oleh peneliti yaitu :

   (a) ***Stop words removal dan word segmentation***
   Pada kegiatan ini peneliti membagi setiap artikel ke dalam bentuk kata untuk menentukan kesamaan makna ataupun bentuk kata. Pada kegiatan ini juga, peneliti mulai menghapus atau mengabaikan kata - kata yang tidak mempengaruhi dari makna sesungguhnya dari setiap artikel.

   (b) ***Stemming***
   Pada kegiatan ini peneliti memotong setiap kata yang terdapat dalam *corpus* untuk menghilangkan awalan dan akhiran pada kata. Proses *stemming* ini digunakan untuk membentuk kesamaan antar fitur-fitur yang sebelumnya hanya berbeda pada awalan dan akhiran.

   (c) ***Feature Selection***
   Pada kegiatan ini peneliti menghitung nilai *chi square* setiap kata yang ada dalam *corpus* untuk menentukan kebergantungan penentuan *genre* terhadap kata tersebut. Pada kegiatan ini kata yang memiliki nilai *chi*

square yang mencukupi dari standar yang ditentukan akan dijadikan sebagai fitur-fitur yang mempengaruhi penentuan genre yang biasa disebut sebagai *bag of words*.

(d) **Membangun *Naive Bayesian Multi-Label Classifier***
Pada kegiatan ini peneliti membentuk model yang dibutuhkan untuk membangun *classifier*. Model yang dibentuk terdiri dari dua yaitu *prior probability* dan *likelihood probabiity*. *Prior probability* dibentuk dengan menghitung *probability* kemunculan per *genre* yang ada dalam *training data* sedangkan *likelihood probability* dibentuk dengan menghitung *probability* kemunculan setiap fitur pada setiap *genre* yang ada dalam *corpus*.

4. **Pengujian**
Pada tahap ini peneliti melakukan pengujian terhadap *classifier* yang telah terbentuk dari tahap implementasi. Pengujian yang dilakukan antara lain adalah melakukan proses *undersampling* terhadap *dataset* yang ada, menentukan sampel *dataset* terbaik, melakukan percobaan terhadap perubahan komposisi persentase *training* dan *testing data* yaitu 75%-25%, 50%-50%, dan 25%-75% serta melakukan percobaan terhadap perubahan *level significant* dari parameter *chi square* yaitu 0.0001, 0.001, 0.01, 0.1 0.25 dan 0 (tanpa *level siginificant*). Pengujian ini dilakukan agar mendapatkan model terbaik untuk *classifier*. Hasil dari setiap pengujian ini akan menampilkan nilai *F1-measure*.

5. **Analisis**
Pada tahap ini peneliti melakukan analisis dari semua tahapan sebelumnya. Analisis yang dilakukan mengenai hasil dari tahap pengujian lalu mencocokkan dengan tujuan penelitian. Analisis yang dilakukan antara lain pengaruh komposisi persentase *training* dan *testing* data terhadap hasil klasifikasi dan pengaruh *level significant* terhadap performansi *classifier* dan jumlah *fitur* yang digunakan sebagai *bag of words*.

6. **Pelaporan**
Pada tahap ini peneliti membuat laporan akhir dari setiap tahap yang bertujuan agar semua hal yang didapatkan dapat terdokumentasi dan diharapkan untuk dilakukannya pengembangan lebih lanjut

1.6 Sistematika Penulisan

Untuk memahami lebih jelas laporan penelitian ini, dilakukan dengan cara mengelompokkan materi menjadi beberapa sub bab dengan sistematika penulisan sebagai berikut:

- **BAB I : PENDAHULUAN**
Bab ini menjelaskan tentang informasi umum yaitu latar belakang penelitian, perumusan masalah, batasan masalah, tujuan penelitian, metodologi penyelesaian masalah, dan sistematika penulisan.

- **BAB II : KAJIAN PUSTAKA**
  Bab ini berisikan teori yang diambil dari beberapa kutipan buku, *paper* dan artikel yang berupa pengertian dan definisi serta persamaan matematika. Bab ini juga menjelaskan konsep *data mining*, *text mining*, *machine learning*, *naive bayesian multi-label classifier*, serta teori lain yang diperlukan dalam proses pembangunan dan analisis pada kasus ini.

- **BAB III : METODOLOGI DAN DESAIN SISTEM**
  Bab ini berisi tentang gambaran umum sistem, cara pengerjaan dan kebutuhan perangkat lunak dan keras. Dalam bab ini juga berisi tentang langkah - langkah pengerjaan dari mulai pengolahan *dataset* sampai pembangunan model *classifier*.

- **BAB IV : PENGUJIAN DAN ANALISIS**
  Bab ini menjelaskan tentang pengujian dan analisis terhadap *classifier* yang telah dibangun. Dalam bab ini peneliti melakukan tahap demi tahap dalam menentukan parameter-parameter yang dibutuhkan untuk membangun model *classifier* yang terbaik dalam kasus ini.

- **BAB V : KESIMPULAN DAN SARAN**
  Bab ini berisi kesimpulan dan saran yang berkaitan dengan analisis dan optimalisasi sistem berdasarkan yang telah diuraikan pada bab-bab sebelumnya.