BAB I

PENDAHULUAN

1.1 Latar Belakang

Text mining adalah ilmu yang mempelajari bagaimana menarik informasi yang menarik, sesuatu yang baru, pola yang belum diketahui sebelumnya atau menemukan kembali informasi tersirat yang berasal dari kumpulan sumber-sumber data teks yang berbeda-beda. *Text mining* mengekstrak informasi atau pola yang berguna dari sumber data teks melalui identifikasi dan eksplorasi dari suatu pola menarik[1]. Sumber data pada *text mining*, berupa sekumpulan dokumen text tidak beraturan yang nantinya akan diolah menjadi sebuah dataset pada tahap *preprocessing* data. Tahapan *preprocessing* data dalam *text mining* adalah tahap yang paling memakan waktu. Dalam tahapan ini ada berbagai proses yang harus dilakukan pada data yang akan dijadikan bahan penelitian. Tahapan data preprocessing tersebut diantaranya : *case folding, tokenizing, stopwords, stemming, lemmatization, normalization,* dan sebagainya[1]. Tahap data *preprocessing* dilakukan untuk mendapatkan dataset yang bersih/ siap digunakan.

Stemming dan *Lemmatization* adalah dua teknik pemrosesan bahasa alami yang penting digunakan dalam *Information Retrieval* (IR) untuk pemrosesan *query* dan dalam *Machine Translation* (MT) untuk mengurangi inkonsistensi data[2]. Keduanya meminimalkan bentuk infleksional, dan terkadang bentuk turunan kata, ke bentuk dasar yang sama. Sebagian besar pekerjaan *stemmer* dan *lemmatization* yang ada didasarkan pada beberapa aturan yang tergantung pada bahasa yang memerlukan pengawasan ahli bahasa, atau beberapa pendekatan probabilistik yang

membutuhkan sejumlah besar *corpus monolingual*, yang keduanya mengembangkan algoritma *stemming* dan *lemmatization* secara mandiri. *Stemming* dan *lemmatization* sekilas memang sama, tapi ada perbedaan mencolok diantara kedua teknik data *preprocessing* ini. *Stemming* merupakan proses untuk memetakan berbagai variasi morfologikal dari kata menjadi bentuk dasar yang sama, dengan menghilangkan semua imbuhan baik yang terdiri dari awalan, sisipan, akhiran dan kombinasi dari awalan dan akhiran pada kata berimbuhan[3]. *Lemmatization* adalah proses yang bertujuan untuk melakukan normalisasi pada teks dengan berdasarkan pada bentuk dasar yang merupakan bentuk lemmanya, lemma merupakan bentuk dasar sebuah kata yang memiliki arti tertentu yang berdasar pada kamus[4]. Singkatnya *stemming* hanya memotong kata berimbuhan, sedangkan *lemmatization* akan mengembalikan kata berimbuhan menjadi kata dasarnya.

Sentiment analysis atau analisis sentimen adalah studi komputasional dari opini-opini orang dengan menggunakan ilmu *text mining*. Analisis sentimen akan mengelompokkan teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut, bisa berupa sentimen positif, maupun negatif[2]. Namun opini - opini orang tersebut tentu tidak semuanya dituliskan dengan menggunakan tata bahasa yang baku. Maka dari itu proses *preprocessing* akan sangat berpengaruh terhadap hasil dari analisis sentimen.

Berdasarkan latar belakang permasalahan diatas, penilitian ini dilakukan untuk mengetahui manakah yang lebih efektif diantara *stemming* dan *lemmatization* dalam tahap *preprocessing* data untuk melakukan analisis sentimen.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk mengetahui bagaimana perbedaan akurasi dari penggunaan *stemming* dan *lemmatization* untuk analisis sentimen ?

1.3 Hipotesis

Dari rumusan masalah tersebut dapat ditarik hipotesis awal berupa:

H_0: Tidak ada perbedaan signifikan tingkat akurasi dari penggunaan *stemming* dan *lemmatization*

H_1: Ada perbedaan signifikan tingkat akurasi dari penggunaan *stemming* dan *lemmatization*

1.4 Batasan Masalah

Batasan masalah dalam penelitian yang akan dilakukan adalah :

1. Dataset yang digunakan diambil dari penyedia data online, Kaggle.com
2. Menekankan pada penggunaan *stemming* dan *lemmatization* dalam data preprocessing
3. Validasi dilakukan pada hasil klasifikasi
4. Menggunakan algoritma *Support Vector Machine* untuk melakukan klasifikasi data yang telah didapatkan

1.5 Maksud dan Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk melihat signifikansi perbedaan tingkat akurasi antara *stemming* dan *lemmatization* untuk analisis sentimen

1.6 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah untuk menentukan manakah yang memiliki tingkat akurasi yang lebih baik diantara *stemming* dan *lemmatization* untuk analisis sentimen, sehingga dapat digunakan untuk acuan penelitian lain kedepannya.

1.7 Metode Penelitian

Dalam penelitian ini ada beberapa tahapan yang perlu dilakukan, diantaranya:

1. Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini diunduh dari kaggle dataset https://www.kaggle.com/columbine/imdb-dataset-sentiment-analysis-in-csv-format#Valid.csv. Data yang digunakan berupa teks yang merupakan sentimen user terhadap film dari IMDB.

2. Metode Analisis Data

Dalam tahap ini ada beberapa proses yang akan dilakukan untuk melakukan analisis data diantaranya:

a. Preprocessing

Preprocessing merupakan salah satu tahapan dalam text mining dimana data mentah yang didapatkan dari twitter akan dibersihkan untuk dijadikan dataset penelitian. Tahap *preprocessing* memiliki beberapa proses, yaitu *case folding, stopwords removal, tokenizing, stemming/ lemmatization,* dan normalisasi. Selanjutnya data yang sudah mengalami *preprocessing* akan diubah menjadi bentuk numerik dengan tahap *term*

weighting. Karena, penelitian ini bertujuan untuk melihat pengaruh dari *stemming* dan *lemmatization* maka dalam tahapan *preprocessing* ini *stemming* dan *lemmatization* akan dibahas lebih mendalam.

Stemming merupakan proses untuk memetakan berbagai variasi morfologikal *dari* kata menjadi bentuk dasar yang sama, dengan menghilangkan semua imbuhan pada kata. Atau sederhananya stemming adalah proses untuk memotong kata berimbuhan. Misal kata-kata seperti "pendidikan", "dididikan", "pendidik" akan dibuang imbuhannya menjadi kata "didik".

Lemmatization adalah proses yang bertujuan untuk melakukan normalisasi pada teks dengan berdasarkan pada bentuk dasar yang merupakan bentuk lemmanya, lemma merupakan bentuk dasar sebuah kata yang memiliki arti tertentu yang berdasar pada kamus.

b. Implementasi Algoritma

Ditahap ini dataset yang sudah diubah menjadi bentuk numerik dalam tahap *term weighting* diklasifikasi menjadi sentimen positif dan negatif menggunakan algoritma *Support Vector Machine* (SVM).

c. Evaluasi Hasil Klasifikasi

Dalam tahap ini data yang telah diklasifikasi menggunakan algoritma *Support Vector Machine* (SVM) kemudian akan dievaluasi performa akurasi, presisi, dan recallnya dengan *Confusion Matrix*.

d. Validasi Klasifikasi

Kemudian pada tahap ini dataset akan divalidasi menggunakan *k-fold cross validation* untuk melihat model mana yang menghasilkan nilai paling optimal untuk analisis sentimen dengan *support vector machine*.

3. Analisis Statistik (Uji Signifikan)

Kemudian melakukan perhitungan statistik komparatif dari hasil akurasi yang didapatkan *stemming* maupun *lemmatization* untuk melihat perbandingan akurasi dari keduanya.

4. Hasil

Terakhir melakukan analisa dari hasil perhitungan komparatif dari keduanya untuk melakukan penarikan kesimpulan apakah ada perbedaan signifikan dari *stemming* dan *lemmatization* dalam proses analisis sentimen.

1.8 Sistematika Penulisan

Penulisan skripsi ini dibagi dalam beberapa bab dengan rincian sebagai berikut:

BAB I PENDAHULUAN

Dalam bab ini berisikan latar belakang, rumusan masalah, batasan masalah, maksud dan tujuan penelitian, manfaat penelitian, metode penelitian, dan sistematika penelitian.

BAB II LANDASAN TEORI

Bab ini menjelaskan tentang tinjauan pustaka dan teori-teori yang akan digunakan didalam penelitian ini.

BAB III METODE PENELITIAN

Dalam bab ini berisi tentang tahapan-tahapan yang akan dilakukan dalam penelitian, misal: metode pengumpulan data, metode analisis data, metode klasifikasi, serta metode uji statistik.

BAB IV ANALISIS DAN PEMBAHASAN

Bab ini berisi tentang analisis dari hasil penelitian yang berupa pengolahan dataset, implementasi algoritma, pengujian data yang telah diolah, dan analisis dari hasil pengujian.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran yang didapat dari hasil proses penelitian