

BAB I

PENDAHULUAN

1.1 Latar Belakang

Platform digital saat ini sudah mengubah cara berinteraksi juga mengubah cara dalam menyampaikan pendapat atau ulasan. Sebagai contoh yaitu media *e-commerce* dengan menuliskan ulasan konsumen terhadap produk yang dibelinya. Dengan partisipasi masyarakat pada platform *e-commerce* ini menghasilkan sejumlah data besar berupa teks subjektif yang berisi sentimen dan pendapat pengguna[1]. Karena setiap ulasan pengguna memiliki peran penting dalam mempengaruhi tingkat penjualan dan minat konsumen terhadap produk yang dibelinya. Oleh karena itu, analisis *semantic similarity* dapat mencari ulasan mana saja yang memiliki maksud dan tujuan yang sama.

Semantic Similarity merupakan salah satu *task* pada *text mining* dalam ranah *Natural Language Processing (NLP)* untuk melihat kemiripan makna pasangan teks dengan memberikan nilai kemiripan antara sepasang kata secara semantik. Banyak riset yang terkait yang telah dilakukan untuk mendapat nilai kemiripan antar teks. Pengukuran kemiripan kata yang telah banyak diketahui biasanya berdasarkan pada *WordNet*[2].

Pada penelitian sebelumnya yang menggunakan metode *Latent Semantic Analysis (LSA)* untuk mencari kemiripan kata agar memperoleh inti dari sebuah

informasi menghasilkan bahwa hasil dari penggunaan metode LSA memperoleh nilai rata-rata akurasi precision, recall, dan f-measure secara berurutan pada compression rate 50% dengan nilai 0.668, 0.743, 0.700 dan 0.690 sedangkan pada compression rate 40% sebesar 0.696, 0.605, 0.642 dan 0.663[3].

Penelitian ini menggunakan metode *Latent Semantic Analysis* (LSA). LSA adalah Teknik dalam pemrosesan Bahasa alami, khususnya dalam vector semantik. LSA menganalisis hubungan antara sekumpulan dokumen dan istilah yang dikandungnya dengan menghasilkan serangkaian konsep yang terkait dengan dokumen[4]. LSA merupakan salah satu metode untuk menemukan hubungan, keterkaitan, dan kemiripan antar dokumen-dokumen, penggalan dari dokumen- dokumen dengan memanfaatkan komputasi statistic untuk menggali dan mempresentasikan konteks yang digunakan sebagai sebuah arti kata untuk sejumlah *corpus* yang besar. *Corpus* adalah kumpulan teks yang memiliki kesamaan subyek atau tema[5].

Adapun algoritma analitik yang akan digunakan dalam penelitian ini yaitu *Topic Modelling*. Ide dasar dari *Topic Modelling* adalah bahwa sebuah topik terdiri dari kata-kata tertentu yang Menyusun topik tersebut, dan dalam satu dokumen memiliki kemungkinan terdiri dari beberapa topik dengan peluang masing-masing. Namun secara pemahaman manusia, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik, distribusi topik per-dokumen dan penggolongan setiap kata pada topik perdokumen merupakan struktur tersembunyi, maka dari itu *Topic Modelling* bertujuan untuk menemukan topik dan kata-kata yang terdapat pada topik tersebut[6].

Berdasarkan penjelasan tersebut maka penulis melakukan sebuah penelitian untuk mendapatkan metode yang paling baik dalam melakukan identifikasi topic modelling pada ulasan produk *e-commerce*. Hasil penelitian yang dilakukan penulis diharapkan dapat menghasilkan system identifikasi ulasan *e-commerce* dengan tingkat akurasi yang tinggi.

1.2 Rumusan Masalah

Rumusan permasalahan yang akan diselesaikan dalam penelitian ini ialah bagaimana cara melakukan identifikasi ulasan yang ditulis oleh konsumen dalam platform *e-commerce* yang memiliki arti dan tujuan sama, sehingga menghasilkan informasi yang lebih ringkas.

1.3 Batasan Masalah

Batasan masalah dari penelitian ini adalah :

- 1 Data yang digunakan adalah data dari Kaggle dengan judul "NLP – Topic Modelling".
- 2 Metode yang digunakan dalam pemodelan topik adalah *Latent Semantic Analysis (LSA)*.
- 3 Library python yang digunakan adalah *gensim*, *nlTK* dan *matplotlib*.
- 4 Dataset yang dipergunakan adalah teruntuk topic modelling.

- 5 Output dari penelitian ini berupa visualisasi diagram grafik dari nilai keakuratan topic ulasan produk.
- 6 Agar penelitian terfokus dan tidak menyimpang, penulis melakukan penelitian seputar Topic modelling dan metode *Latent Semantic Analysis*.

1.4 Maksud dan Tujuan Penelitian

Tujuan dari penelitian ini adalah melakukan identifikasi *Topic Modelling* untuk mengetahui topik yang memiliki arti dan tujuan sama yang sering diulas, sehingga dapat menghasilkan informasi yang lebih ringkas.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

- 1 Mempermudah dalam menerima informasi yang ringkas dari ulasan yang diberikan konsumen.
- 2 Mengetahui kemampuan algoritma *Latent Semantic Analysis* dalam melakukan identifikasi pada pendekatan *Topic Modelling*.

1.6 Metode Penelitian

1.6.1 Metode Pengumpulan Data

Data yang digunakan pada penelitian ini dapat dari *website* Kaggle (www.kaggle.com) . Kaggle adalah sebuah platform dan situs yang

menyediakan berbagai dataset dari berbagai macam perusahaan yang dapat digunakan untuk melakukan identifikasi *Topic Modelling*. Data yang digunakan menggunakan data yang berjudul “NLP – *Topic Modelling*”.

Data ini berisi ulasan konsumen terhadap produk makanan yang dibelinya disitus penjualan *online* atau *e-commerce*. Dataset ini dirilis pada tahun 2017 dan terakhir di *update* pada tahun 2018. Dalam dataset “NLP - *Topic Modelling*” terdapat atribut *Id, Product Id, User Id, Profile Name, Score, Time, Text*. Dari banyak atribut yang ada, yang akan banyak digunakan adalah atribut *Text*.

1.6.2 Metode Analisis

Pada tahap ini melakukan analisis mengenai bagaimana metode *Latent Semantic Analysis* (LSA) bekerja untuk mengidentifikasi pemodelan topik pada ulasan produk agar informasi yang didapat lebih ringkas.

1.6.3 Metode Perancangan

Dalam metode perancangan terdapat beberapa tahapan yang akan dilakukan. Setelah mendapatkan dataset yang merupakan tahap awalan, selanjutnya melakukan pengolahan terhadap dataset yang didapat dengan melakukan data cleaning, tokenization, stopword removal dan stemming. Setelah pengolahan data selesai dilanjutkan dengan melakukan tahap *Topic Modelling* data untuk mendapatkan variable tersembunyi, yaitu sebuah topik. Setelah mendapatkan data topik dilanjutkan dengan tahap *Semantic Similarity*

menggunakan algoritma *Latent Semantic Analysis* untuk mendapatkan nilai keakuratan dari identifikasi kesamaan kata dari ulasan produk yang memiliki arti dan tujuan sama.

1.6.4 Metode Pengujian

Pada tahap ini dilakukan pengujian pada dataset yang menjadi bahan utama untuk identifikasi kesamaan kata menggunakan *Latent Semantic Analysis* (LSA). Dengan menghitung representasi nilai keakuratan dengan menggunakan metode LSA pada pemodelan topik.

1.7 Sistem Penulisan

Sistematika penulisan dibuat untuk mempermudah penulis dalam penyusunan skripsi. Adapun sistematika penulisan ini dikelompokkan kedalam beberapa bab. Setiap bab diuraikan sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini menjelaskan tentang dasar penelitian, yang berisi latar belakang, rumusan masalah, Batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian dan sistematika penulisan.

BAB II METODE PENELITIAN

Pada bab ini berisi tinjauan Pustaka yang mirip dengan

penelitian ini. Pada bab ini juga berisi tentang landasan-landasan teori yang mendukung dalam penelitian ini.

BAB III METODE PENELITIAN

Pada bab ini berisi tentang alur dari penelitian yang berupa perancangan perangkat dan bahan apa saja yang akan digunakan.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang hasil dari tahapan penelitian yang dilakukan secara menyeluruh termasuk hasil dari pengujian.

BAB V PENUTUP

Bagian terakhir dari penelitian yang berisi tentang kesimpulan dan saran untuk memperbaiki kekurangan yang ada pada penelitian ini.