BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam mengimbangi keberagaman informasi yang diinginkan manusia, media massa dihadirkan sebagai jalan yang menunjukkan bahwa arus globalisasi sedang berjalan dan akan siap untuk memenuhi keinginan manusia akan informasi. Manusia sebagai khalayak yang menikmati media massa juga harus bisa memilih informasi yang sesuai dengan kebutuhannya, diharuskan untuk lebih teliti untuk menerima pesan media agar tidak salah dalam menerima informasi yang disajikan media itu sendiri (Gunawan, 2017). Menurut kompas pada tahun 2019, sejak berita-berita hoaks marak, Kompas.com secara reguler menjadi bagian dari media massa yang berusaha memverifikasi dan memvalidasi setiap berita hoaks atau fakta yang beredar di masyarakat. Kerja jurnalistik Kompas.com secara otomatis menempatkan diri sebagai *fact-checker* dari setiap simpang-siur berita yang ada.

Pemberitaan dalam media masa tertulis (*online* dan *offline*) sering dikategorikan sesuai dengan tema dari isi berita untuk memudahkan pencarian dan pemahaman konteks. Kategorisasi (klasifikasi) ini secara umum dilakukan oleh penulis berita pada saat berita tersebut ditulis. Proses klasifikasi ini mempunyai panduan pelaksanaan yang baik, seperti berbasis frekuensi kata (*word density*), namun pada pelaksanaannya sering kali dilakukan secara subjektif oleh penulis. Klasifikasi secara subjektif dapat mengakibatkan efek negatif yaitu klasifikasi

berita yang kurang tepat dan tambahan proses bisnis dari penulis berita seperti yang saat ini terjadi di media masa kompas.com.

Klasifikasi berita di kompas.com dilakukan oleh penulis berita sesuai dengan opini pribadi. Hasil dari klasifikasi tersebut sulit untuk di evaluasi sehingga penelusuran tingkat akurasi dan pertanggung jawabannya. Efek-efek negatif tersebut dapat dikurangi dengan cara mengurangi tingkat subjektifitas melalui proses klasifikasi berita secara otomatis. Klasifikasi secara otomatis akan mengikuti proses algoritma yang terstruktur sehingga tingkat akurasinya dapat diukur secara objektif dan dapat ditelusuri pertanggung jawabannya (Soelistio, 2019).

Penelitian ini dilaksanakan berdasarkan penelitian internal dosen prodi Informatika, Sistem Informasi dan Jurnalistik terkait masalah pada sistem kategori berita harian kompas. Pada penelitian ini, penulis akan membangun sebuah model pengelompokan berita menggunakan dua metode pengelompokan. Model akan dirancang dan dibangun dengan algoritma K-Means dan Gaussian Mixture Model. Menurut Yuan & Yang pada tahun 2019, Algoritma K–Means yang merupakan *hard clustering* memiliki banyak keunggulan seperti perhitungan matematika yang sederhana, konvergensi yang cepat dan implementasi yang mudah. K-Means juga merupakan salah satu algoritma pengelompokan yang paling populer dan hal pertama yang biasanya diterapkan untuk mendapatkan gambaran tentang struktur *dataset*.

Adapun penelitian yang dilakukan oleh Joshi & Kaur (2013) mengenai studi banding teknik clustering dalam data mining menyimpulkan Algoritma K-Means memiliki keuntungan besar dalam melakukan pengelompokan terhadap *dataset*

besar dan kinerjanya meningkat dengan meningkatnya jumlah *cluster*. Tidak seperti K-Means, Algoritma Gaussian Mixture Model yang merupakan *soft clustering* memakan waktu kerja lebih lama, dikarenakan komputasi yang lebih kompleks. Gaussian Mixture Model memperhitungkan nilai varians dan mengembalikan probabilitas bahwa suatu titik data menjadi milik masing-masing *cluster*. Gaussian Mixture Model dapat mengatasi kelemahan pengelompokan, pada ambiguitas data yang tumpang tindih dalam Algoritma K-Means (Celeux dkk, 2010).

Dalam penelitian Reddy dkk (2019), dilakukan uji coba perbandingan performa algoritma K-Means dan Gaussian Mixture Model dalam mengelompokkan data gambar. *Dataset* yang digunakan dalam penelitian ini sebanyak 60.000 gambar. Hasil uji coba menunjukan Gaussian Mixture Model memiliki tingkat akurasi yang lebih baik yaitu 99% dibandingan dengan K-Means yang hanya memiliki tingkat akurasi sebesar 96%.

Pada penelitian Husni dkk (2015), dilakukan uji coba clusterisasi dokumen web (berita) Bahasa Indonesia Menggunakan Algoritma K-Means. Dalam penelitian ini, Berita dikelompokan secara otomatis dengan kelompok-kelompok berita yang memiliki kesesuaian sebanyak dua sampai dengan sepuluh *cluster*. *Dataset* yang digunakan dalam penelitian ini sebanyak 300 berita. Hasil uji coba menyimpulkan dokumen berita berhasil dikelompokkan secara otomatis sesuai dengan kesamaan berita dengan tingkat akurasi tertinggi sebesar 61%. Akurasi yang belum sempurna dapat ditingkatkan dengan memperbaiki teknik *pre-processing* dan melakukan uji coba analisis menggunakan metode lain.

Adapun penelitian serupa yang dilakukan oleh Wibisono & Khodra (2005) mengenai Clustering Berita Berbahasa Indonesia. Penelitian menggunakan 4.718 data berita yang diambil dari situs www.kompas.com pada bulan Juni – November 2005. Kategori berita yang diberikan kompas adalah Metro, Otomotif, Kesehatan, Olahraga, Teknologi, dan Gaya Hidup. Dalam penelitian ini tahap *pre-processing* menggunakan 329 *stopword* dan Porter Stemmer yang disesuaikan dengan aturan bahasa Indonesia. Pada Tahap *Clustering* berita dikelompokkan menggunakan algoritma K-Means. Hasil uji coba penelitian menunjukan tingkat akurasi tertinggi pengelompokan berita sebesar 56%. Peneliti menyimpulkan bahwa *Clustering* pada berita berbahasa Indonesia memiliki potensi yang besar untuk dikembangkan lebih lanjut dan juga diperlukanya penelitian lanjutan untuk meningkatkan kualitas pengelompokan.

Penelitian-penelitian tersebut menunjukan bahwa *Text Clustering* memiliki potensi yang besar untuk dikembangkan lebih lanjut. Menurut Husni dkk (2015), penyempurnaan teknik *pre-processing* dan uji coba dalam menggunakan metode-metode lain dapat meningkatkan akurasi pada *clustering*. Berdasarkan penelitian yang sudah ada sebelumnya, Kelebihan pada masing-masing algoritma K-Means dan Gaussian Mixture Model mampu untuk melakukan clustering dan berpotensi untuk dikembangkan lebih lanjut. Maka, dalam penelitian ini akan dikembangkan serta dibandingkan performa algoritma K-Means dan Gaussian Mixture Model dengan menggunakan sumber berita di layanan RSS Kompas (2020).

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan sebelumnya, Rumusan masalah yang akan dijawab pada penelitian ini mencakup:

1. Bagaimana mengimplementasikan pengelompokan berita dengan menggunakan algoritma K-Means dan Gaussian Mixture Model?
2. Bagaimana performa dan kecepatan eksekusi algoritma K-Means dan Gaussian Mixture Model dalam melakukan pengelompokan berita?

1.3 Batasan Masalah

Batasan masalah pada penelitian ini telah ditentukan bahwa ruang lingkup model hanya mendalami pada masalah pokok, yaitu:

1. Total data yang akan digunakan adalah 10 ribu data berita, dengan kelompok berita: travel, news, bola, edukasi, global, tren, money, health, sains, otomotif, properti, tekno, lifestyle, hype, skola, advertorial (RSS Kompas, 2020).
2. Penghitungan bobot kata (*Term Weighting*) pada masing-masing algoritma menggunakan metode Term Frequency – Inverse Document Frequency (TF-IDF).
3. Tahapan proses *stemming* bahasa Indonesia menggunakan *library* python Sastrawi.
4. Kedua algortima *clustering* menggunakan *library* python scikit-learn.
5. Evaluasi performa dari masing–masing algoritma menggunakan metode *silhouette coefficient*.

1.4 Tujuan Penelitian

Adapun tujuan dalam penelitian ini, sebagai berikut:

1. Melakukan pengelompokan berita di situs kompas.com dengan menggunakan algoritma K-Means dan Gaussian Mixture Model.
2. Membandingkan performa algoritma K-Means dan Gaussian Mixture Model dalam melakukan pengelompokan berita di kompas.com.

1.5 Manfaat Penelitian

Adapun manfaat dalam penelitian ini, sebagai berikut:

1. Sebagai rujukan untuk para pengembang dalam membangun sistem text clustering, sistem tagging berita dan penelitian sejenis lainya.
2. Sebagai rujukan untuk para pengembang dalam menentukan algoritma clustering yang efektif untuk digunakan.

1.6 Sistem Penulisan

Sistematika penulisan laporan penelitian disusun dan dibagi atas lima bab sebagai berikut.

1. BAB I PENDAHULUAN

   Bab ini berisi latar belakang permasalahan, rumusan masalah, batasan masalah, manfaat penelitian, dan sistematika penulisan laporan.

2. BAB II LANDASAN TEORI

   Bab ini mendeskripsikan tentang teori-teori dan konsep dasar yang berkaitan dengan penelitian yang dilakukan. Teori maupun konsep dasar yang

digunakan antara lain Berita, Media Massa, *Information Retrieval, Pre-Processing, Case Folding*, *Filtering*, *Stemming*, *Term Weighting*, K–Means, GMM, PCA, & *Silhouette Coefficient*.

3. BAB III METODOLOGI PENELITIAN DAN PERANCANGAN APLIKASI

   Bab ini berisi tentang metode penelitian yang digunakan dan perancangan aplikasi berupa flowchart.

4. BAB IV IMPLEMENTASI DAN UJI COBA

   Bab ini memuat implementasi dan hasil dari uji coba aplikasi.

5. BAB V KESIMPULAN DAN SARAN

   Bab ini berisi simpulan dari hasil penelitian dan saran untuk pengembangan selanjutnya.