BAB I

PENDAHULUAN

1.1 Latar Belakang Penelitian

Informasi menjadi kebutuhan yang vital bagi masyarakat saat ini. Kebutuhan informasi semakin meningkat setiap tahunnya baik di bidang politik, hukum, ekonomi, bisnis, sampai hiburan. Informasi bisa didapat dari majalah, surat kabar, buku, website, media sosial dan lain-lain. Namun saat ini media sosial menjadi sumber informasi yang populer dan menjadi alternatif baik untuk media pencari dan penyebar informasi.

Salah satu media sosial yang populer saat ini adalah *Twitter*. *Twitter* adalah *microblogging* dan situs jejaring sosial dunia yang populer, memungkinkan penggunanya untuk mem-*posting* pesan pendek atau *tweet* sampai dengan 140 karakter. Saat ini jumlah pengguna aktif *twitter* mencapai 320 juta pengguna dengan rata-rata jumlah *tweet* lebih dari 300 juta *tweet* per harinya (Edwards, 2016). *Twitter* hadir sebagai sarana komunikasi untuk bertukar informasi mengenai berbagai peristiwa di dunia nyata. Pesan-pesan singkat pada *twitter* secara umum mencerminkan berbagai peristiwa yang dialami pengguna secara *real-time* (Becker, Naaman, & Gravano, 2011).

Twitter menyediakan sumber informasi yang besar dan mudah didapat. Sifat informasinya mengalir secara *stream* dan *up-to-date*. Maka dari itu *twitter* bisa digunakan sebagai sumber *real-time* untuk identifikasi peristiwa di dunia nyata. Topik pembicaraan di *twitter* sangat beragam, sehingga dibutuhkan pemroses *tweet* untuk memperoleh informasi yang berharga.

Salah satu fitur yang menarik pada *twitter* adalah *trending topics. Trending topics* adalah kumpulan topik tertentu yang banyak dibicarakan dalam *tweet* pengguna. *Trending topics* berfungsi untuk mengetahui percakapan yang sedang dibahas saat ini dan membantu pengguna untuk selalu *update* tentang kejadian terbaru dan menemukan masalah utama dari masyarakat.

Trending topics memiliki peran penting dalam menemukan berita atau kejadian terhangat dan aktual dengan cepat. Sehingga *trending topics* telah

menarik minat besar tidak hanya di kalangan pengguna sendiri tetapi juga di kalangan konsumen informasi lain seperti wartawan, pengembang aplikasi *real-time*, dan peneliti media sosial.

Banyak penelitian yang telah dilakukan untuk menggali dan mencari *trending topics* dengan berbagai macam teknik yang ditawarkan. Seperti pada penelitian yang dilakukan oleh Zubiaga, Spina, Martinez, & Fresno (2013) mengenai proses klasifikasi secara *real-time* terhadap data *tweet*. Selain itu penelitian serupa dilakukan oleh Becker, Naaman, & Gravano (2011) mengenai proses *clustering* terhadap data *stream twitter* kemudian dilakukan proses klasifikasi untuk membedakan *cluster event* dan *non-event*.

Penelitian terhadap *trending topics* juga dilakukan oleh Aiello, Petkos, & Martin (2013) yaitu dilakukan perbandingan terhadap enam metode deteksi topik pada tiga dataset *Twitter* terkait dengan peritiwa besar. Penelitian oleh Sahdev & Kabra (2013), mendeteksi *trending topics* dengan pendekatan graf jaringan sosial untuk menentukan perilaku individu dengan menghubungkan interaksi individu dengan individu lainnya. Penelitian oleh Berhandus (2010), mendeteksi dan mengidentifikasi *trending topics* dari data *stream*. Penelitian oleh Lau, Collier, & Baldwin (2012) yang menyajikan metodologi berbasis pemodelan baru untuk melacak peristiwa yang muncul pada *microblog Twitter*. Penelitian Lu & Yang (2012) melakukan analisis tren pada topik berita di *twitter*, yang meliputi prediksi tren dan analisis penyebab terjadinya perubahan tren. Penelitian Miller, Vodrahalli, & Lee (2015) menawarkan sebuah model untuk memperkirakan *k* topik *hastag twitter* terpopuler pada interval waktu tertentu atau pada suatu periode. Kemudian penelitian oleh Mathioudakis & Koudas (2010) menyajikan sebuah sistem yang melakukan deteksi tren pada *stream Twitter* dan melakukan analisis tren. Penelitian oleh Kim, Kim, Rho, & Hwang (2013) mengusulkan skema baru untuk mendeteksi tren dan kata kunci dari aliran data *Twitter*.

Mendeteksi *trending topics* bukanlah hal yang mudah, butuh pendekatan khusus untuk menganalisis aliran data *tweet* yang datang secara terus-menerus dari jutaan pengguna *twitter*. Data yang digunakan sangatlah besar sehingga dibutuhkan media penyimpanan yang besar, pengolahan data yang cukup lama, dan algoritma yang efisien untuk analisis data. Maka dari itu, dibutuhkan

pendekatan secara *streaming* sebagai strategi yang efisien untuk menemukan informasi dari data yang besar. Selain itu, pendekatan secara *streaming* akan menghasilkan *trending topics* secara *real-time* dan *up-to-date.*

Salah satu teknik yang digunakan dalam menemukan *trending topics* adalah metode *clustering*. *Clustering* adalah suatu proses menguji kumpulan "titik" dan mengelompokkan titik-titik tersebut ke dalam "*clusters"* berdasarkan ukuran jarak (Leskovec, Rajaraman, & Ullman, 2014). Metode *clustering* digunakan pada penelitian (Rosa, Shah, & Lin, 2011), pengelompokkan *tweet* dilakukan secara otomatis dengan dua kali percobaan menggunakan algoritma *Latent Dirichlet Allocation* (LDA) dan algoritma *K-means*. Hasil dari kedua *clustering* menunjukkan bahwa *clustering* dengan algoritma *k-means* memiliki kualitas *cluster* lebih baik dibanding dengan algoritma LDA. Evaluasi *cluster* dilakukan dengan menggunakan metode *Purity*.

Clustering dapat digunakan untuk menganalisis topik pada *tweet* dengan mengelompokkan secara otomatis *tweet* yang memiliki kesamaan. *Clustering* pada teks atau dokumen berbeda dengan *clustering* pada data terstruktur. Pada *text clustering* diperlukan algoritma pengelompokan yang dapat menangani data berdimensi tinggi. *K-means* adalah salah satu algoritma *unsupervised learning* yang paling sederhana yang dapat memecahkan masalah *clustering* (MacQueen, 1967). *K-means* merupakan metode *clustering* yang sangat terkenal dan banyak digunakan di berbagai bidang karena sederhana, mudah diimplementasikan, memiliki kemampuan untuk melakukan *cluster* dengan data yang besar dan mampu menangani data *outlier*.

Algoritma *K-means* dapat dijalankan dengan dua cara yaitu *batch mode* dan *online mode* (Aaron & Rishe, 2014). *Online mode* disebut juga dengan *sequential mode* merupakan kebalikan dari *batch* yang akan menyimpan semua data, sedangkan pada *online mode* hanya menyimpan sejumlah data seperti *cluster centers*. *Sequential k-means* (Aaron & Rishe, 2014) termasuk ke dalam algoritma *online clustering* yang digunakan berdasarkan pada kebutuhan komputasi yang ketat dan kebutuhan operasi pada data *online*. Algoritma *online clustering* digunakan untuk mengelompokkan data *stream* yang datang secara terus-menerus dan tidak terbatas. Ketika data baru tiba algoritma harus memasukkannya ke

dalam salah satu *cluster* yang ada atau membuka *cluster* baru dengan data tunggal (Beringer & Hullermeier, 2006).

Penelitian tentang *online clustering* pada data *stream* pernah dilakukan sebelumnya oleh (Beringer & Hullermeier, 2006). Penelitian tersebut bertujuan untuk mengelompokkan data *stream* dan menghasilkan struktur *cluster* yang *up-to-date*. Teknik *clustering* yang digunakan merupakan versi *online* dari algoritma pengelompokan klasik *k-means* yang efisien.

Penelitian mengenai *trending topics* merupakan topik yang menarik untuk diteliti. Oleh karena itu, penulis ingin melakukan penelitian untuk menemukan *trending topics* menggunakan data *stream twitter* dengan pendekatan *sequential k-means*. Karakteristik topik pada penelitian ini yaitu menggunakan kata bebas yang terlepas dari *hastags*. Aplikasi *trending topics* akan dibuat dengan menggunakan bahasa R, yaitu bahasa pemrograman dan lingkungan perangkat lunak yang menyediakan lingkungan interaktif yang kuat untuk komputasi ilmiah, analisis data, visualisasi, pemodelan, *machine learning*, komputasi kinerja tinggi, statistik dan lainnya (Ihaka & Gentleman, 1996). Alasan menggunakan bahasa R adalah penelitian ini menggunakan *package* yang tersedia di *Comprehensive R Archive Network* (CRAN). Aplikasi akan menampilkan lima topik teratas yang akan menjadi *trending topics* dalam bentuk histogram.

1.2 Rumusan Masalah Penelitian

Berdasarkan latar belakang di atas, permasalahan yang timbul yaitu:

1. Bagaimana merancang model dan strategi dalam mendeteksi *trending topics* dari data *stream twitter* dengan pendekatan *sequential k-means*?
2. Bagaimana membangun aplikasi deteksi *trending topics* dari data *stream twitter* dengan menggunakan bahasa R?
3. Bagaimana kualitas *cluster* yang dihasilkan dari data teks *stream* dengan menggunakan algoritma *sequential k-means*?
4. Bagaimana hasil analisis perbandingan penelitian dengan penelitian lain yang serupa?

1.3 Batasan Masalah Penelitian

Untuk memfokuskan penelitian, ditetapkan beberapa batasan masalah yaitu sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah *tweet stream* berbahasa Inggris dikarenakan *package* pendukung yang digunakan pada aplikasi dikhususkan untuk teks berbahasa Inggris.
2. *Tweet* yang digunakan dalam penelitian ini adalah *tweet* dengan lokasi di kota New York.
3. Adanya loncatan waktu saat pengambilan data *stream* karena tertunda oleh proses pengolahan data.
4. Karakteristik topik adalah kata bebas yang terlepas dari *hastags* dan kata yang bersifat kata benda (*noun*).
5. Sistem hanya akan menampilkan lima topik teratas berdasarkan jumlah *tweet*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Merancang model dan strategi dalam mendeteksi *trending topics* dari data *stream twitter* dengan pendekatan *sequential k-means*.
2. Membangun aplikasi deteksi *trending topics* dari data *stream twitter* dengan menggunakan bahasa R.
3. Menguji kualitas *cluster* yang dihasilkan dari data teks *stream* dengan menggunakan algoritma *sequential k-means*.
4. Melakukan analisis perbandingan penelitian dengan penelitian lain yang serupa.

1.5 Manfaat Penelitian

Setelah melakukan penelitian ini, ada pengharapan dari penulis atas terciptanya manfaat yang diuraikan sebagai berikut:

Bagi Penulis

1. Produk dari hasil penelitian ini dapat dijadikan penulis sebagai acuan untuk melanjutkan penelitian-penelitian selanjutnya.

2. Kajian teoritis yang dilakukan pada penelitian ini bisa meningkatkan pemahaman penulis dari segi teori maupun praktis sehingga dapat membantu proses penulisan karya ilmiah yang akan datang.

Bagi Dunia

1. Survey terhadap situasi terkini dapat dilakukan secara gratis.
2. Pendekatan dengan menggunakan data *streaming* merupakan pendekatan baru dalam pencarian *Trending Topics*.
3. Tidak adanya masalah dengan memori dalam data *gathering* pada data *stream* karena penelitian ini menggunakan konsep *single-pass*.

1.6 Sistematika Penulisan

Adapun sistematika penulisan mengacu pada Pedoman Penulisan Karya Ilmiah Universitas Pendidikan Indonesia Tahun 2015 yang terurai sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas latar belakang penelitian "Deteksi *Trending Topics* dari Data *Stream Twitter* dengan Pendekatan *Sequential K-Means*" serta dibahas pula rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian bagi penulis serta bagi dunia dan sistematika penulisan skripsi.

BAB II KAJIAN PUSTAKA

Bagian kajian pustaka dalam skripsi memberikan konteks yang jelas terhadap topik atau permasalah yang dikaji dalam penelitian. Bagian ini berisi hasil studi literatur tentang Media Sosial sebagai Sumber Informasi, *Twitter*, *Trending Topics*, *Data Stream*, *Text Mining*, *Text Clustering* (Algoritma *K-Means*, Algoritma *Sequential K-means*, Variasi *Sequential K-Means*) Evaluasi *Cluster*, Deteksi Topik dan Bahasa R. Studi literatur didapat dari sumber yang dapat dipertanggung jawabkan.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai metodologi dalam "Deteksi *Trending Topics* dari Data *Stream Twitter* dengan Pendekatan *Sequential K-Means*" dan pembangunan aplikasi menggunakan bahasa R. Metodologi penelitian yang

dilakukan berisi Desain Penelitian (studi literatur, pengumpulan data *stream twitter*, desain model dan strategi deteksi *trending topics*, pengembangan perangkat lunak, rancangan eksperimen, analisis hasil dan penarikan kesimpulan), Metode Pengembangan Perangkat Lunak serta Alat dan Bahan Penelitian yang digunakan.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi uraian mengenai hasil penelitian dan pembahasan terhadap penelitian yang dilakukan. Untuk menguji model yang telah dirancang, eksperimen dilakukan dalam tiga skenario dengan jumlah data yang berbeda serta waktu pengambilan data yang berbeda, dan nilai parameter untuk fungsi yang digunakan berbeda untuk setiap skenario. Pengujian terhadap kualitas *cluster* dilakukan dengan menggunakan metode *Dunn Index*.

BAB V KESIMPULAN DAN SARAN

Bab ini menjelaskan mengenai kesimpulan secara keseluruhan dari penelitian dan saran untuk penelitian selanjutnya yang berkaitan tentang pencarian *trending topics twitter*.