BAB I
PENDAHULUAN

1.1. Latar Belakang

Penggunaan komputer untuk menyelesaikan masalah telah dilakukan untuk segala bidang pekerjaan. Hal ini karena komputasi dianggap lebih cepat dalam menyelesaikan masalah dibandingkan penyelesaian secara manual. Seiring dengan hal tersebut, dituntut proses komputasi yang semakin cepat. Maka untuk meningkatkan kecepatan proses komputasi dapat ditempuh dengan dua cara, yaitu peningkatan kecepatan perangkat keras dan peningkatan kecepatan perangkat lunak. "Peningkatan kecepatan/performa perangkat keras *(microprocessor)* mengalami perkembangan 2 faktor setiap 18 bulan, atau dikatakan meningkat 60% setiap tahun*"* (Parhami, 2012). Peningkatan performa kinerja suatu komputasi pada akhirnya berkembang dan mulai memanfaatkan *multiple processor* dan dalam hal ini melakukan pemrosesan paralel.

Pemrosesan paralel menjadi sebuah pilihan setelah pemrosesan sekuensial mengalami berbagai masalah dan keterbatasan. Hal ini disebabkan karena kecepatan pemrosesan sekuensial belum mencukupi kebutuhan bidang sains dan rekayasa akan kecepatan komputasi yang tinggi. Pemrograman mempergunakan pemrosesan paralel untuk membuat progaram yang menyelesaikan suatu masalah dengan waktu eksekusi yang lebih kecil. Singkatnya dengan pemrosesan paralel maka komputasi akan dilakukan secara bersama-sama oleh beberapa komputer.

Perubahan arsitektur komputer menjadi *multiprocessor* memang bisa membuat lebih banyak proses bisa dikerjakan sekaligus, namun tetap tidak bisa meningkatkan kecepatan masing-masing proses (David dan Wen-Mei, 2010). Peningkatan kecepatan setiap proses bisa dicapai melalui peningkatan perangkat lunak. Kecepatan perangkat lunak sangat ditentukan oleh algoritmanya,

dengan adanya komputer *multiprocessor,* maka algoritma yang lebih cepat dapat dirancang dengan memanfaatkan paralel proses komputasi.

Komputer *multiprocessor* yang masih jarang dan mahal harganya menyebabkan algoritma paralel yang ada sukar diimplementasikan, selain itu aplikasi dari komputasi paralel juga mengalami kesulitan untuk dijalankan. Pembuatan aplikasi tersebut bertujuan untuk menunjukkan peningkatan kecepatan yang diperoleh dari paralelisasi algoritma sekuensial. Maka, untuk mengatasinya dapat dilakukan dengan merancang mesin paralel semu. David dan Wen-mei (2010) mengatakan bahwa mesin paralel semu yang dirancang dapat dilakukan dengan beberapa cara, dengan *message passing interface,* dengan suatu jaringan komputer, atau dengan menggunakan *graphic card* GPU.

Penelitian yang dilakukan oleh Masyhudi (2012) *general purpose* (GP) *graphics processing unit* (GPU) yang dikembangkan oleh perusahaan NVIDIA tahun 2006, saat ini *graphics card* memiliki kemampuan untuk mengolah proses-proses komputasi dan tidak hanya untuk pengolahan citra saja. GPU menawarkan daya komputasi yang jauh lebih cepat dengan biaya yang sangat rendah. NVIDIA juga mengembangkan bahasa pemrograman CUDA *(compute unified device architecture)* untuk membuat program yang dapat dijalankan pada CPU dan GPU. Bagian-bagian serial dari program akan dijalankan pada CPU dan bagian-bagian paralel akan dijalankan pada GPU. Meskipun CUDA dan perkembangan terkait lainnya mempercepat penggunaan GPU untuk aplikasi tujuan umum yang memparalelkan komputasi, namun beberapa tantangan masih tetap dalam pemrograman GPU. Oleh karena itu, jelas diinginkan pengembangan dan penelitian lebih lagi untuk dapat menggunakan program GPU dalam menghasilkan antarmuka tingkat yang lebih tinggi (Garland dkk, 2008)

Salah satu pengolahan data pada jumlah yang besar adalah untuk mengolah data yang diperoleh dari internet. Sebagai contoh untuk *micro-blogging* pada saat ini menjadi alat komunikasi paling populer dan digemari oleh para pengguna internet. Pengguna berbagai opini tentang berbagai aspek kehidupan sehari-hari dengan sifat postingan yang *real-time,* sehingga *micro-blogging* menjadi situs yang kaya akan data tekstual (Pak dan Paroubek, 2010). *Micro-blogging* juga

menjadi sumber dari beragam jenis informasi yang dapat dipakai sebagai bahan untuk melakukan suatu revolusi dalam suatu negara, seperti yang terjadi dinegara Arab (Akaichi, 2013). Twitter merupakan nama website yang menyediakan layanan *micro-blogging.* Sebagai salah satu sosial media *on-line,* twitter menjadi penyedia dan penyebar informasi yang sangat cepat. Wijaya dkk. (2013) mengatakan bahwa informasi yang beredar sangat bebas dan beragam seperti berita, pertanyaan, opini, komentar, kritik baik yang bersifat positif maupun negatif.

Ketersediaan data yang cukup besar dapat dimanfaatkan untuk *text mining* yang mengacu pada proses mengambil informasi yang berkualitas tinggi dari teks. Informasi yang berkualitas tinggi biasanya diperoleh melalui peramalan pola dan kecenderungan melalui sarana seperti pembelajaran pola satistik (Saraswati, 2011). *Text mining* menjadi topik yang menarik untuk diteliti dan diolah pada saat ini, karena ketersediaan dokumen teks yang banyak dan kemudahan dalam memperoleh data, dalam upaya memberikan informasi yang lebih baik. Pengorganisaian data tekstual bagi pengguna, para peneliti telah menyelidiki masalah tentang kategorisasi teks secara otomatis. Beberapa penelitian mulai dilakukan untuk mengetahui sentimen dari penulis yang terkandung pada teks yang dituliskan, dan penelitian ini disebut dengan analisis sentimen.

Penelitian yang dilakukan oleh Aliandu (2012) bahwa analisis sentimen adalah riset kumputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual. Yusuf dan Santika (2011) mengatakan bahwa analisis sentimen merupakan proses memahami, mengekstrak, dan mengolah data tekstual secara otomatis untuk mendapatkan informasi. Analisis sentimen mencoba untuk mengumpulkan pendapat secara keseluruhan terhadap komentar-komentar, misalnya perusahaan *micro-blogs* berusaha mempelajari reaksi pengguna untuk mendapatkan rasa umum terhadap produk mereka. Passonneau dkk (2011), twitter sebagai salah satu *micro-blogging* membangun model untuk mengklasifikasikan "Tweet" menjadi sentimen positif, negatif dan netral. Dengan fasilitas twitter API yang dimilikinya, twitter mampu menyediakan data terkini atas tweet yang disampaikan oleh penggunanya.

Proses analisis sentimen dilakukan dengan cara memanfaatkan teori klasifikasi. Klasifikasi suatu teks dapat dibagi menjadi kelas positif, negatif atau netral. Klasifikasi dapat dilakukan dengan barbagai cara, baik dengan metode SVM, K-Means, Naïve Bayes, dll. Metode Naïve Bayes yang sering disebut dengan *Naïve Bayes Classifiers* (NBC). Penelitian sebelumnya yang diacu pada penelitian ini adalah yang dilakukan oleh Aliandu (2012) yang menggunakan NBC dalam penelitiannya untuk melakukan analisis sentimen tweet. NBC memiliki kelebihan dibanding dengan algoritma metode lain, karena algoritma yang digunakan sederhana tetapi memiliki akurasi yang tinggi. Algoritma NBC yang sederhana dengan kecepatannya yang tinggi dalam pelatihan dan klasifikasi membuat algoritma ini menarik untuk digunakan sebagai salah satu metode klasifikasi.

Pada penelitian yang dilakukan oleh Aliandu (2012) untuk mengetahui hasil akurasi pada uji akurasi aplikasi dengan menggunakan metode Holdout, proses dilakukan mulai dari tanggal 15-12-2011 pukul 23:43:18 dan selesai pada tanggal 16-12-2011 pukul 03:24:09 membutuhkan waktu 73.149 detik = 1.219,15 menit = 20,3191667 jam. Penelitian ini menunjukkan bahwa melakukan proses analisis sentimen tweet ternyata membutuhkan waktu yang sangat lama, tentu saja hal ini sangat memprihatinkan dari sisi komputasi dan efisiensi. Berdasarkan hasil penelitian tersebut, maka penelitian ini dilakukan untuk mengatasi masalah bagaimana membuat suatu aplikasi yang dapat memanfaatkan GPU untuk memparalelkan proses klasifikasi pada proses analisis sentimen dengan tujuan memperoleh komputasi yang lebih cepat.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, maka yang menjadi rumusan masalah dalam penelitian ini adalah bagaimana menerapkan paralelisasi *naïve bayes classifiers* untuk analisis sentimen tweet menggunakan *Graphics Processing Unit* (GPU)?

1.3. Batasan Masalah

Berdasarkan perumusan masalah, maka dalam penelitian ini dilakukan pembatasan masalah sebagai berikut:

1. Data yang digunakan adalah data komentar aplikasi dalam bahasa Indonesia yang diperoleh dari API twitter.
2. Metode klasifikasi yang digunakan untuk melakukan klasifikasi komentar-komentar adalah *Naïve Bayes Classifiers*
3. Ulasan tekstual aplikasi yang diklasifikasikan menjadi tiga buah polaritas sentimen yaitu: positif, negatif dan netral.
4. Proses paralel yang dilakukan adalah pada proses pengujian/test data bersih.
5. Bahasa pemrograman yang digunakan adalah Java dan CUDA untuk GPU

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah mengembangkan suatu metodologi paralellisasi *naïve bayes classifiers* untuk analisis sentimen tweet menggunakan GPU sehingga proses yang dihasilkan dapat lebih cepat.

1.5. Keaslian Penelitian

Sejauh pengamatan penulis, penelitian mengenai analisis sentimen yang dilakukan dengan menggunakan *naïve bayes classifiers* sudah pernah dilakukan sebelumnya. Namun penulis belum pernah menjumpai penelitian tentang paralelisasi metode *Naïve Bayes Classifiers* yang diimplementasikan untuk analisis sentimen menggunakan dan diimplementasikan pada GPU.

1.6. Manfaat Penelitian

Sistem yang dihasilkan dapat digunakan untuk melakukan analisis sentimen, serta memperoleh implementasi yang efisien dari model paralel *naïve bayes classifiers*.

1.7. Metode Penelitian

Metode penelitian dalam mengembangkan sistem ini dibagi menjadi beberapa tahap, yaitu:

1. **Pengumpulan data dari Informasi**

a. Data dari API Tweeter

Pada tahap ini dengan studi kasus analisis sentimen pada data Tweeter API, dan mulai melakukan pengumpulan data yang akan digunakan sebagai data training dan pengujian pada sistem nantinya.

b. Studi Literatur

Pada tahap selanjutnya setelah ditemukan beberapa contoh kasus untuk analisis sentimen maka akan dilakukan studi literatur untuk mempelajari teori analisis sentimen dan teori pemrograman paralel untuk algoritma *naïve bayes classifiers* menggunakan CUDA, metode yang digunakan dan aplikasi yang diterapkan dengan membaca buku-buku teori, tesis, jurnal penelitian, prosiding dan sumber-sumber yang diperoleh melalui internet.

2. **Analisis dan rancangan sistem**

Pada tahap ini dilakukan analisis kebutuhan sistem. Sistem yang dibuat harus dapat melakukan analisis sentimen untuk *user input* yang berupa *query.* Proses pembelajaran dan klasifikasi pada pengolahan data, sedangkan data komentar tentang *query* yang diinputkan oleh *user* diperoleh dari API twitter.
Selanjutnya dengan pengetahuan dan referensi yang ada maka akan dilakukan perancangan sistem yang akan dikembangkan.

3. **Implementasi**

Pada tahap ini dilakukan penerapan dari perancangan yang dilakukan dengan melakukan pengolahan data, pembuatan *baseline,* dan menggunakan metode *machine learning* untuk analisis sentimen. Pengembangan sistem yang dibangun menggunakan bahasa CUDA untuk pemrograman pada GPU dengan metode *naïve bayes classifiers.*

4. **Pengujian dan evaluasi**

Pada tahapan ini dilakukan untuk mengetahui kinerja *classifier* untuk menghitung nilai akurasi hasil klasifikasi yang diprediksi benar dan yang tidak

benar. Data pengujian diperoleh dari data komentar yang dimasukkan secara acak untuk opini positif dan opini negatif. Hasil pengujian akan dievaluasi tingkat akurasinya jika diperoleh nilai akurasi yang cukup tinggi maka kinerja *classifier* baik. Selain mengevaluasi tingkat akurasi, pada tahap ini juga akan dilakukan analisis kecepatan dalam proses klasifikasi dan pengujian sistem. Dengan paralel algoritma yang dilakukan akan dibandingkan kecepatan prosesnya dengan algoritma klasifikasi yang dijalankan konvensional atau satu saja. Jika pemanfaatan GPU terbukti mempercepat proses maka penelitian ini mencapai tujuan dalam pelaksanaannya.

Proses evaluasi kinerja *classifier* menggunakan pendekatan *k-fold cross-validation*. Pendekatan *k-fold cross-validation* mengeneralisasi pendekatan *cross-validation* dengan mensegmentasi data ke dalam *k* partisi berukuran sama. Dalam pendekatan *cross-validation*, setiap *record* digunakan beberapa kali dalam jumlah yang sama untuk *training* dan tepat satu kali untuk *testing*. Selama proses, salah satu dari partisi dipilih untuk *testing*, sedangkan sisanya digunakan untuk *training*. Prosedur ini diulangi *k* kali sedemikian sehingga setiap partisi digunakan untuk testing tepat satu kali. Total *error* ditentukan dengan menjumlahkan *error* untuk semua *k* proses tersebut. Pada penelitian ini pengujian menggunakan *k-fold cross-validation*, dengan *k*=3 sesuai dengan penelitian Pang dkk (2002) dan penelitian Nur dan Santika (2011)yang menggunakan *3-folds cross-validation* karena dataset yang digunakan cukup besar dan proses *training classifier* yang membutuhkan waktu yang lama. *Confusion matrix* digunakan untuk menghitung nilai akurasi sedangkan evaluasi untuk proses paralel *Naïve Bayes* yang dilakukan dengan menghitung efektivitas dari suatu proses paralel (Parhami, 2002).

1.8. Sistematika Penulisan

Sistematika yang digunakan dalam penulisan materi dalam tesis ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini memberikan informasi mengenai latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, metodologi penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini akan memaparkan beberapa referensi dari penelitian-penelitian sebelumnya yang terkait dengan analisis sentimen twitter, dan pemrograman paralel. Selain itu juga dipaparkan perbedaan antara penelitian sebelumnya yang sejenis sebagai bahan pembanding dengan penelitian yang dilakukan.

BAB III LANDASAN TEORI

Bab ini akan menjelaskan tentang konsep dan teori sentimen analisis, *naïve bayes classifier, feature selection,* pemrosesan paralel*, graphic processing unit* (GPU) dan bahasa pemrograman CUDA yang digunakan dalam penelitian ini.

BAB IV ANALISIS DAN RANCANGAN

Bab ini berisi tentang gambaran umum *naïve bayes clssifiers*, rincian perancangan paralelisasi algoritma *naïve bayes classifier* untuk analisis sentimen yang meliputi analisis algoritma, rancangan proses, dan rancangan antarmuka.

BAB V IMPLEMENTASI

Bab ini akan membahas mengenai implementasi sesuai analisis dan rancangan yang dijelaskan pada bab IV.

BAB VI HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi hasil pengujian performansi algoritma *naïve bayes classifier* yang dilakukan secara paralel untuk analisis sentimen.

BAB VII KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran yang diambil berdasarkan hasil penelitian yang sudah dilakukan.