1. Pendahuluan

1.1 Latar Belakang

Penggunaan ponsel semakin bertambah seiring dengan berkembangnya teknologi informasi dan hal tersebut didukung dengan semakin banyaknya fasilitas media komunikasi yang diberikan seperti munculnya berbagai aplikasi *messenger* yaitu *whatsapp, line, bbm,* dan masih banyak lagi lainnya. Namun, walaupun sudah banyak aplikasi *messenger* yang ditawarkan, SMS (*short message text*) tetap menjadi salah satu media komunikasi yang menjadi pilihan karena kemudahan penggunaannya untuk berbagai kalangan pengguna, tarifnya yang murah, aman dan dapat didokumentasikan.

SMS adalah sebuah media komunikasi berbentuk teks yang mengijinkan pengguna ponsel untuk saling berbagi teks pendek (biasanya kurang dari 160 karakter 7-bit) [10]. Seiring dengan penggunaannya yang semakin meluas dan popularitasnya sebagai media komunikasi terpenting, banyak pihak yang memanfaatkan hal tersebut untuk kepentingan komersil seperti sebagai media iklan bahkan penipuan. Menurunnya tarif SMS menjadi salah satu penyebab juga semakin meningkatnya SMS *spam,* seperti di Cina tarif untuk SMS kurang dari $0.001 [10]. Jumlah SMS *junk* atau SMS *spam* semakin bertambah setiap harinya dan berdasarkan *Korea Information Security* (KISA), jumlah SMS *junk* ini melebihi email *spam*. Sebagai contohnya, pengguna ponsel di US mendapatkan 1,1 milyar SMS *spam* dan pengguna Cina juga menerima 8.29 milyar SMS *spam* dalam seminggu [13].

Salah satu solusi yang dapat dilakukan terhadap permasalahan diatas adalah melakukan *filtering* SMS dengan klasifikasi teks. Beberapa teknik yang populer untuk klasifikasi teks diantaranya *decision trees, Naive Bayes, rule induction, neural network, nearest neighbors,* dan *Support Vector Machine.* Namun dalam klasifikasi SMS ini berbeda dengan klasifikasi pada teks dokumen biasa atau email dikarenakan teks pada SMS sangat pendek (maksimal 160 7-bit karakter), banyak terdapat teks yang disingkat, dan cenderung tidak formal [10]. SMS yang terlalu pendek ini menimbulkan pertanyaan lain "apakah fitur yang digunakan cukup untuk membedakan antara SMS *spam* dengan *non-spam*?". Bahkan kini jenis SMS semakin bervariasi, sehingga dibutuhkan teknik lain untuk penambahan fitur yang dapat membedakan antara SMS *spam* dengan *non-spam.* Namun, dari setiap variasi SMS yang ada tetap memiliki pola yang serupa khususnya untuk sms *spam,* hal tersebut dapat menjadi landasan untuk digunakan teknik dengan melibatkan kemunculan kata-kata yang muncul bersamaan sebagai fitur tambahan untuk membedakan sms *spam* dan *non-spam*.

Dalam penelitian ini akan digunakan kolaborasi dua buah metode yaitu *Naive Bayes Classifier* dan *FP-Growth Algorithm Frequent Itemset*. *Naive Bayes* dianggap sebagai salah satu *learning algorithm* yang sangat efektif dan penting untuk *machine learning* dalam *information retrieval.* Selain itu berdasarkan paper yang diacu [2] menyatakan bahwa dengan menerapkan *minimum support* yang ditentukan pengguna, dapat meningkatkan akurasi dibandingkan dengan hanya menggunakan *Naive Bayes* saja. Karena dengan *minimum support* didapatkan

frequent itemset sebagai fitur tambahan, sehingga tidak hanya setiap kata yang dianggap *mutually independent*, tetapi juga kata yang *frequent* sebagai kata yang *single, independent* dan *mutually exclusive* [2], maka mampu meningkatkan nilai peluang dan menyebabkan sistem lebih tepat dalam klasifikasi. Dalam paper acuan digunakan *Apriori Algorithm* dalam mendapatkan *frequent itemset*, namun pada penelitian ini dipilih menggunakan *FP-Growth Algorithm* yang memiliki kemampuan lebih baik dibandingkan *Apriori Algorithm* [4]. *FP-Growth* merupakan algoritma *mining frequent pattern* yang memiliki performansi yang baik dan efisien karena tidak membutuhkan pembangkitan kandidat *frequent pattern* [4].

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, permasalahan yang diteliti dalam tugas akhir ini adalah sebagai berikut.

1. Bagaimana mengimplementasikan metode *Naive Bayes Classifier* dan *FP-Growth Algorithm Frequent Itemset* dalam melakukan klasifikasi SMS *spam* dan SMS *ham* untuk digunakan dalam SMS *Filtering* ?
2. Bagaimana pengaruh penerapan metode *FP-Growth Algorithm* terhadap akurasi dan performansi untuk SMS *Filtering* ?
3. Bagaimana tingkat akurasi yang diperoleh dari hasil kolaborasi penerapan metode *Naive Bayes Classifier* dan *FP-Growth Algorithm Frequent Itemset* untuk SMS *Filtering ?*

1.3 Tujuan

Tujuan yang diharapkan oleh penulis dalam pengerjaan tugas akhir ini adalah sebagai berikut.

1. Mengimplementasikan metode *Naive Bayes Classifier* dan *FP-Growth Algorithm Frequent Itemset* dalam melakukan klasifikasi SMS *spam* dan SMS *ham* untuk digunakan dalam SMS *Filtering,*
2. Menganalisis pengaruh penerapan metode *FP-Growth Algorithm* terhadap akurasi dan performansi untuk SMS *Filtering,*
3. Mengetahui tingkat akurasi yang diperoleh dari hasil kolaborasi penerapan metode *Naive Bayes Classifier* dan *FP-Growth Algorithm Frequent Itemset* untuk SMS *Filtering.*

1.4 Batasan Masalah

Batasan masalah untuk tugas akhir ini adalah :

1. Data yang digunakan merupakan data berbahasa inggris, dikarenakan data berbahasa Indonesia sedang dalam proses pengumpulan. Data berasal dari SMS *Corpus* : SMS *Spam Collection v.1* dan SMS *Spam Corpus v.0.1 Big,*
2. Program yang dibangun bersifat *offline,*
3. Data diproses secara *offline,* yaitu tidak terhubung ke internet maupun jaringan lainnya,
4. Data hanya dalam bentuk *text* murni atau *plain text,*

5. Program yang dibangun hanya dapat menangani input teks berbahasa Inggris.
6. Data yang digunakan maksimal 160 karakter yaitu data *text* SMS.

1.5 Metodologi Penyelesaian Masalah

Metodologi yang digunakan untuk menyelesaikan tugas akhir ini adalah sebagai berikut.

1. Studi Literatur
   Tahapan pengumpulan materi-materi, informasi dan referensi terkait dengan permasalahan yang akan dibahas, mencakup teori, metode dan algoritma yang berkaitan dengan permasalahan. Dalam kasus ini mencari literatur terkait *text mining*, SMS *spam*, SMS *classification, naive bayes classifier*, *frequent itemset* atau *frequent pattern, FP-Growth.*
2. Pengambilan dan Analisis Data
   Pada tahap ini, penulis melakukan pengambilan data dari *SMS Corpus* : *SMS Spam Collection v.1* dan SMS *Spam Corpus v.0.1 Big*. Selanjutnya dilakukan analisis terhadap data yang ada untuk mengenali pola data untuk selanjutnya dapat menentukan proses yang tepat untuk dilakukan terhadap data.
3. Perancangan Sistem
   Setelah memahami metode, algoritma dan data yang akan digunakan, langkah selanjutnya adalah merancang sistem yang akan mengimplementasikan algoritma yang dipilih. Rancangan sistem digambarkan menggunakan *flowchart* dari setiap tahapan agar dapat lebih memahami alur yang terjadi di dalamnya. Selain itu, melakukan spesifikasi terhadap fungsionalitas sistem.
4. Implementasi
   Pada tahap ini dilakukan pengimplementasian terhadap sistem berdasarkan rancangan yang telah dibuat untuk mengolah data yang telah diambil. Implementasi pada penelitian tugas akhir ini dibangun pada aplikasi berbasis Java (seperti NetBeans IDE).
5. Analisis Hasil
   Melakukan analisis berdasarkan hasil klasifikasi dan melihat tingkat akurasi yang didapat berdasarkan klasifikasi dengan metode yang digunakan (*Naive Bayes Classifier* dan *FP-Growth Frequent Itemset*).
6. Pembuatan Laporan Tugas Akhir
   Membuat laporan untuk mendokumentasikan hasil penelitian lengkap dengan lampiran-lampiran yang dapat mendukung penelitian Tugas Akhir ini.