BAB II

TINJAUAN PUSTAKA

Beberapa peneliti yang melakukan penelitian menganggap *text mining* menjadi sangat penting karena kemudahan untuk mendapatkan data elektronik dari berbagai macam sumber, karena itu diperlukan klasifikasi yang tepat dan pengalian pengetahuan dari sumber – sumber ini dapat dijadikan sebagai bahan penelitian yang penting (Baharudin, Lee and Khan, 2010). Beberapa peneliti telah mempelajari teknik klasifikasi untuk teks menggunakan menggunakan *machine learning* (Ikonomakis, 2005). Klasifikasi teks bisa diterapkan di berbagai bidang seperti deteksi bahasa dari suatu teks, pengarsipan dokumen, klasifikasi halaman web dan pembelajaran otomatis (Bijalwan, Kumar, Kumari and Pascual, 2014).

Karena panjang pesan dari pesan teks kecil maka kata kunci yang dapat digunakan untuk melakukan klasifikasi juga menjadi lebih kecil daripada *email* selain karena pesan teks juga tidak memiliki *header* serta penuh dengan singkatan dan bahasa yang tidak formal mengakibatkan turunnya performa dari algoritma untuk melakukan *spam filtering* pada pesan teks (Shirani-mehr, 2012). Karena sifat dari SMS yang hanya berjumlah 160 karakter, maka diperlukan suatu metode untuk meningkatkan akurasi dari klasifikasi teks. Peneliti berhasil menggunakan metode *decision tree* untuk meningkatkan akurasi dalam melakukan klasifikasi teks, tetapi terdapat masalah lain seperti melakukan pergantian semua kata – kata pendek yang mungkin untuk kata yang diberikan secara dinamis oleh kata aslinya adalah suatu isu yang harus dibahas (Padhiyar, 2013).

Penelitian untuk meningkatkan akurasi dari metode TF-IDF dalam melakukan klasifikasi teks juga telah dilakukan dengan menggunakan *feature word.* Walaupun telah memberikan hasil, tetapi permasalahan yang lain adalah menentukan nilai kepercayaan yang tepat untuk *corpus* yang berbeda (Zhang, Gong and Wang, 2005).

Pada penelitian yang lain membahas tentang adanya faktor selain *term frequency*, seperti untuk bobot istilah lokal yang diukur dalam satu dokumen seperti TF, ditemukan bahwa istilah dengan frekuensi yang lebih tinggi dan dekat dengan distribusi hipo-dispersi harus diberikan bobot yang lebih tinggi dari satu dengan frekuensi yang lebih rendah dan mendekati dengan distribusi intensif. Di sisi lain, untuk *weight term* global yang dihargai di seluruh koleksi dokumen seperti *Inverse Document Frequency* (IDF) atau frekuensi dokumen terbalik, itu juga menemukan bahwa, dalam koleksi tersebut, istilah dengan frekuensi yang lebih tinggi dan distribusi dengan jenis hipo-dispersi biasanya berisi sedikit informasi.

Karena TF - IDF hanya memerlukan frekuensi *term* ke dalam pertimbangan, maka TF-IDF juga memiliki kelemahan sebagai berikut . Pertama, algoritma TF menghitung *term weight* hanya berdasarkan pada frekuensi mereka . Artinya, *weight* istilah positif berkorelasi dengan frekuensi mereka. Sebenarnya , istilah dengan tinggi frekuensi hanya intensif didistribusikan di bagian dokumen. Hal tersebut cenderung untuk mewakili isi dari bagian bukan seluruh dokumen. Namun , algoritma TF akan menetapkan *term weight* yang lebih tinggi untuk hal

tersebut dan itu tidak cukup jika hanya mempertimbangkan frekuensi *term* ketika menghitung *term weight*.

Kedua, makna intuitif algoritma IDF adalah bahwa hal yang jarang terjadi selama koleksi dokumen adalah berharga. Pentingnya setiap istilah diasumsikan berbanding terbalik dengan jumlah dokumen yang memiliki istilah itu muncul. Namun istilah yang terjadi secara luas dalam koleksi dokumen tetapi intensif muncul dalam beberapa dokumen lebih mungkin merupakan topik kategori dokumen dan signifikan untuk klasifikasi teks . Namun , skenario seperti ini benar-benar diabaikan oleh IDF. Algoritma IDF akan menetapkan *term weight* rendah untuk hal tersebut. Itu tidak cukup untuk hanya mempertimbangkan frekuensi *term* ketika mengukur *term weight*.

Ketiga, istilah kosong dan hal fungsi , termasuk penghubung , preposisi , beberapa keterangan, istilah tambahan, partikel modal, biasanya ada dengan frekuensi tinggi. Hal ini menyebabkan tugas berat yang tidak akurat untuk hal tersebut. Meskipun istilah berhenti tabel selalu digunakan , masalah ini tidak bisa sepenuhnya diselesaikan (Xia and Chai, 2011).

Dalam klasifikasi himpunan dataset dibagi menjadi pelatihan dan uji dataset. Dataset training digunakan dalam membangun model klasifikasi, sedangkan record data uji digunakan dalam memvalidasi model. Model ini kemudian digunakan untuk mengklasifikasikan dan memprediksi dataset baru yang berbeda dari kedua pelatihan dan dataset uji. Algoritma pembelajaran terawasi (seperti klasifikasi) lebih disukai untuk algoritma pembelajaran tidak terawasi (seperti pengelompokan) karena pengetahuan awal tentang label kelas pada dataset

membuat pilihan fitur / atribut mudah dan ini menyebabkan baiknya akurasi prediksi / klasifikasi (Padhiyar, 2013). Klasifikasi teks mengunakan *machine learning* LVQ telah digunakan untuk melakukan klasifikasi teks berbahasa Arab Ada langkah-langkah yang berbeda yang digunakan untuk mengukur keberhasilan klasifikasi yaitu akurasi, presisi, ingat , F - ukuran dan waktu . Parameter lima algoritma LVQ ini telah dipilih secara empiris dengan sedikit peningkatan dan penurunan nilai mereka dan analisis output (Azara, Mohammed, Fatayer, Tamer, El-Halees, 2012).

Tantangan dari klasifikasi untuk teks pendek seperti SMS adalah kurangnya jumlah data yang akhirnya menjadi tantangan tersendiri dalam algoritma pembelajaran dalam praktek nyata. Dalam rangka untuk mendapatkan kinerja yang lebih baik daripada classifier individu dapat melakukan , classifier ensemble pembelajaran berbasis membuat keputusan akhir dengan menggabungkan multi-hasil dari beberapa pengklasifikasi individu (Liu and Wang, 2010).

Jumlah teks *feature* yang terlalu banyak tidak hanya akan mengakibatkan lamanya proses komputasi tetapi juga menurunkan akurasi dari klasifikasi. Konsekuensi yang di hadapi membuat pemilihan *feature* menjadi penting untuk mempercepat proses komputasi dan meningkatkan akurasi, profil dari fitur yang dipilih dengan metode seleksi fitur adalah salah satu indikator yang baik untuk efektivitas metode tersebut. Jika fitur khas ditugaskan skor tinggi dengan metode seleksi fitur , akurasi klasifikasi diperoleh fitur tersebut kemungkinan besar akan lebih tinggi . Sebaliknya , jika fitur yang tidak relevan ditugaskan skor tinggi dengan metode seleksi fitur , akurasi yang diperoleh fitur tersebut akan

terdegradasi. (Uysal and Gunal, 2012). Teks SMS memerlukan perlakuan yang berbeda untuk perwakilan *feature* yang diambil, tipe *feature* yang ada dan bahkan pengklasifikasi yang berbeda terhadap pesan email yang lebih panjang untuk mendapatkan performa yang baik. Di anjurkan untuk mengkonfigurasi terlebih dahulu mesin pembelajaran yang akan digunakan sesuai dengan aspek yang ada (Healy, Delany and Zamolotskikh, 2005).

Filter otomatis untuk SMS spam merupakan tantangan yang masih dihadapi hingga sekarang. Ada 3 masalah utama yang menghalangi perkembangan algoritma yang dapat digunakan pada bidang ini yang pertama adalah kurangnya jumlah dataset yang dapat digunakan, jumlah feature yang sangat kecil yang dapat diambil, dan teks yang berisi singkatan ataupun idiom (Almeida, Hidalgo and Yamakami, 2011).

Kategorisasi teks Kroasia menggunakan Kata Non - Standar (NSW) sebagai fitur. Hasil penelitian menunjukkan bahwa bentuk kata non - standar dapat digunakan sebagai fitur untuk representasi teks dalam kategorisasi teks. Terkait dengan taksonomi NSW adalah bentuk yang sesuai latar belakang pengetahuan atas dasar NSW dapat secara otomatis diambil dari teks. Telah terbukti bahwa NSW membawa informasi yang cukup tentang sifat teks, yang cocok untuk klasifikasi lebih lanjut. Dengan pendekatan ini mungkin untuk secara signifikan mengurangi dimensi dari fitur vektor (space) dan pada saat yang sama mencapai hasil kategorisasi teks yang baik . Menggunakan fitur NSW , vektor fitur telah beberapa kali lebih kecil dimensi daripada dimensi yang asli dan cara ini mengurangi masalah data yang jarang (Beliga and Martinčić-Ipšić, 2014).

Sejumlah besar studi klasifikasi teks memanfaatkan *Bag of Words* (BoW) model yang mewakili dokumen teks di mana urutan yang tepat dari kata-kata , atau istilah , dalam dokumen diabaikan tetapi jumlah kejadian jangka dianggap . Setiap istilah yang berbeda dalam koleksi dokumen akibatnya merupakan fitur individu . Syarat ditugaskan bobot tertentu yang mewakili kepentingan mereka dalam sebuah dokumen yang diberikan. Hasil dari penelitian dan eksperimen yang telah dilakukan adalah kombinasi dari *Bag of Words* (BoW) dan struktur *feature* sering kali lebih baik daripada hanya menggunakan BoW saja dalam melakukan klasifikasi teks singkat seperti SMS (Uysal, Gunal, Ergin and Gunal, 2012).

.