BAB 1
PENDAHULUAN

1.1 Latar belakang

Pada zaman yang modern ini banyak orang yang menggunakan media *e-news* sebagai sarana membaca berita bagi pengguna internet. *E-News* itu sendiri adalah suatu media yang terhubung oleh internet yang digunakan untuk proses pemberitaan informasi dalam dunia maya. Jenis berita yang disajikan pada media tersebut biasanya terdapat bermacam-macam seperti olahraga, politik, dan hiburan. Jaringan ilmu pengetahuan, bisnis, berita dan sosial banyak mengalami peningkatan yang signifikan dan menjadi sangat bermanfaat bila digunakan dalam lingkup yang benar dan sesuai (Renata dan Maharani, 2012).

Banyak pengguna internet pada dunia maya, bersamaan dengan pemilik akun Twitter di berbagai negara semakin meningkat dari tahun-ketahunnya. Twitter adalah salah satu dari berbagai jenis jejaring sosial yang membantu penggunanya untuk memperoleh berbagai informasi melalui unggahan yang disebut kicauan (*tweets)* dengan jumlah maksimal karakter yaitu 280. Jika seseorang telah menjadi pengguna Twitter maka dapat memiliki hak akses untuk mendapatkan informasi *tweets* dari suatu akun tertentu, pengguna harus menjadi pengikut (*follower*) akun tersebut terlebih dahulu (Kwak et al.,2010). Selain itu, apabila terdapat suatu akun terpilih oleh Twitter yang memberikan info tentang politik, namun sesekali akun tersebut mengomentari bahkan memberikan *tweets* tentang hiburan, maka oleh Twitter masih dianggap informasi tentang politik.

Sampai saat ini terdapat banyak sekali *tweets* yang terus bermunculan dan tersebar oleh pengguna Twitter di seluruh negara. *Tweets* yang ada pada beranda Twitter tercampur menjadi satu dan tidak terkelompokkan berdasarkaan jenis beritanya yaitu olahraga, kesehatan, politik, ekonomi, teknologi, wisata dan lain sebagainya. Tidak adanya pengkategorian *tweets* membuat pengguna Twitter kesulitan untuk membaca dan memilahnya berdasarkan informasi yang diinginkannya. Contohnya jika terdapat pengguna yang ingin mencari informasi tentang politik, maka pengguna tersebut harus mencari satu persatu *tweets* berita yang berkaitan dengan politik. Terkadang pengguna jika ingin mendapatkan jenis informasi yang sama, harus menjelajah dalam satu akun Twitter informasi atau berita yang memiliki konten satu jenis. Oleh karena itu, dibutuhkan sebuah sistem yang dapat melakukan klasifikasi *tweets* sesuai dengan jenis informasinya.

Teknik *text mining* beberapa tahun terakhir menjadi sangat populer yang dilatarbelakangi semakin banyaknya jumlah teks digital yang luas dan tidak terstruktur, oleh karena itu perlu dilakukan analisis isi dari konten tersebut dengan cara yang fleksibel (Hearst, 1999). Teknik yang paling menonjol dari *text mining* adalah klasifikasi teks yang menggunakan pembelajaran mesin, yaitu sistem yang memiliki prediksi otomatis mengenai satu atau lebih kategori yang sesuai untuk teks tidak terstruktur dengan bahasa alami (e.g., Inggris, Spanyol, etc.). Klasifikasi

teks merupakan penelitian utama yang banyak digunakan oleh banyak peneliti dari aplikasi komersial.

Salah satu metode klasifikasian teks yang sering digunakan adalah *K-Nearest Neighbor*. Penelitian yang dilakukan oleh Ramadhan dan Zeniarja (2016) menunjukkan bahwa *K-Nearest Neighbor* memiliki performa yang bagus dalam klasifikasi teks dengan akurasi 80%. Selain itu penelitian lainnya yang menggunakan *K-Nearest Neighbor* adalah Nurjanah dan Perdana (2017) yang memiliki akurasi 82%. Dengan melihat hasil tersebut dapat dipahami bahwa penggunaan metode *K-Nearest Neighbor* memang baik dan bagus digunakan pada pengklasifikasian teks.

Metode klasifikasi teks secara umum masih memiliki kekurangan jika diterapkan pada *short-text* seperti *Twitter* (Liu dan Fan, 2012). Kelemahan yang dimiliki oleh *short text* adalah adanya ambiguitas dan sedikitnya kata pada teks (Tang dan Wang, 2017). Penyebab dari ambiguitas adalah teks tersebut bersifat pendek dan berisi hanya beberapa kata yang mungkin akan ditemukan kesamaan kata dengan lebih dari satu kategori jika diklasifikasikan. Selain itu, *short text* hanya berisi sedikit kata yang terkadang jarang digunakan pada data latih, yang pada akhirnya sistem tidak dapat mengklasifikasikan kategori mana yang tepat. Teknik yang baik guna membantu permasalahan tersebut adalah dengan menambahkan fitur atau *query* baru atau yang disebut *query expansion*. Caranya adalah mengoptimasi pesan atau teks singkat (*short text*) dengan menambahkan beberapa kata yang memiliki kesamaan atau kedekatan secara semantik. Dengan menggunakan *query expansion* maka perbendaharaan kata sebelumnya akan bertambah lebih banyak lagi.

Terdapat juga banyak penelitian yang menggunakan *query expansion.* Penggunaan ini cukup populer diterapkan pada penelitian. Penelitian yang dilakukan oleh Soekarno (2016) menggunakan metode tersebut guna memperbaiki tingkat akurasi, dari hasil asli klasifikasi yang hanya 80% meningkat menjadi 82% setelah ditambahkan metode tersebut. Selain itu juga melalui penelitian Roi dan Ali (2016) dengan menggunakan metode tambahan *query expansion* dapat menghasilkan 96% dari yang sebelumnya 93% ketika masih belum ditambahkan metode tersebut. Dengan demikian metode tersebut terbukti mampu untuk menambahkan tingkat akurasi yang dilakukan pada hasil klasifikasi.

Salah satu cara melakukan *query expansion* adalah menggunakan informasi eksternal (*Unlabeled Background Knowledge*) seperti Wikipedia, Wordnet, dan dokumen berita (Zelikovitz dan Hirsh, 2000). *Query expansion* bisa dilakukan dengan menambahkan beberapa kata dari informasi eksternal ke dalam data teks yang akan diklasifikasi. Kata-kata yang ditambahkan adalah kata-kata kedekatan secara semantik, atau yang memiliki keterkaitan satu kata dengan yang lainnya. Kedekatan semantik tersebut dapat dihitung dengan model semantik terdistribusi (MST) dan dimasukkan ke dalam sebuah kamus yang besar sebagai sumber pengetahuan eksternal untuk sistem. MST adalah sumber semantik eksternal yang dibangkitkan otomatis dan hal tersebut dilatarbelakangi dengan

asumsi bahwa kata kata yang secara semantik mempunyai arti yang sama akan muncul dalam konteks kata-kata yang sama (Yudi,2016).

Berdasarkan latar belakang tersebut, maka diusulkan penelitian yang berjudul **"PENERAPAN KLASIFIKASI TWEETS PADA BERITA TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR DAN QUERY EXPANSION BERBASIS DISTRIBUTIONAL SEMANTIC"**. Diharapkan penelitian ini dapat membantu pengguna media sosial membaca suatu konten berita sesuai kebutuhannya dengan lebih mudah.

1.2 Rumusan masalah

Terkait latar belakang yang telah dijelaskan di atas maka didapatkan rumusan masalah di bawah ini, yaitu:

1. Bagaimana menerapkan metode K-Nearest Neighbor pada klasifikasi teks *tweets* pengguna *Twitter*?
2. Bagaimana pengaruh nilai *k* pada klasifikasi tweets menggunakan KNN?
3. Bagaimana pengaruh penambahan fitur (*query expansion*) berbasis *Distributional Semantic* pada klasifikasi *tweets* menggunakan KNN?

1.3 Tujuan

Terdapat beberapa tujuan penulis terhadap dibuatnya penelitian ini. Adapun tujuan-tujuan tersebut adalah sebagai berikut.

1. Implementasi klasifikasi teks menggunakan metode K-Nearest Neighbor.
2. Mengetahui pengaruh nilai *k* dan mendapatkan nilai *k* terbaik pada klasifikasi *tweets* menggunakan KNN.
3. Mengetahui pengaruh *query expansion* berbasis *Distributional Semantic* dan mendapatkan parameter terbaik pada klasifikasi *tweets* menggunakan KNN.

1.4 Manfaat

Penelitian yang dilakukan ini diharapkan dapat memberikan sebuah manfaat, antara lain yaitu:

Manfaat secara umum

1. Membantu pengguna Twitter guna memperoleh informasi secara mudah
2. Membantu pengguna Twitter dalam memfilter informasi yang dibutuhkan sesuai dengan konten dan kategorinya.
3. Pengguna dapat memperoleh informasi yang akurat sesuai dengan yang diinginkannya.

Manfaat bagi universitas/instansi

1. Memperkenalkan produk baru dari penelitian mahasiswa terhadap masyarakat.
2. Mengukur keberhasilan mahasiswa dari proses perkuliahan yang diberikan terhadap mahasiswa.

Manfaat bagi Mahasiswa

1. Mahasiswa dapat mengetahui pemahaman mengenai metode klasifikasi dan *query expansion* berbasis *distributional semantic*.
2. Dapat menerapkan ilmu yang diperoleh dari jurusan Informatika Universitas Brawijaya kepada masyarakat.

1.5 Batasan masalah

Untuk memberikan ruang lingkup yang jelas terhadap suatu penelitian maka dibuatlah batasan-batasan penelitian sebagai berikut.

1. Data *tweets* yang digunakan pada Twitter hanya berbahasa indonesia.
2. Data latih dan uji diambil dari *tweets* akun detik dan kompas
3. Metode klasifikasi yang digunakan adalah K-Nearest Neighbor
4. Tidak membandingkan hasil klasifikasi yang diperoleh dengan penelitian klasifikasi lainnya karena memiliki perbedaaan metode dan data.
5. Pada ekspansi kata (*query expansion*), kata yang ditambahkan berdasarkan kamus kedekatan kata hanya pada data uji. Data latih tidak dilakukan ekspansi kata.
6. Klasifikasi dibagi menjadi beberapa kategori, yaitu travel, otomotif, olahraga, teknologi, hiburan, makanan, ekonomi, dan kesehatan.
7. Implementasi yang diterapkan pada penelitian ini tidak dapat menangani teks singkatan dan kata tidak baku.

1.5 Sistematika pembahasan

Sistem penulisan penelitian ini menggunakan kerangka pembahasan yang berupa:

BAB 1 Pendahuluan

Berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat, sistematika penulisan.

BAB 2 Landasan Teori

Menguraikan teori terkait acuan yang dijadikan dalam penyusunan penelitian ini seperti yang ada pada materi di mata kuliah pendukung untuk penelitian ini.

BAB 3 Metode Penelitian

Bab ini membahas metode dan langkah-langkah yang digunakan dalam penyusunan penelitian ini mengenai perancangan dan pembangunan aplikasi konsultasi bimbingan skripsi.

BAB 4 Perancangan Sistem

Menggambar dan mendeskripsikan desain global dari sistem yang akan dibuat.

BAB 5 Implementasi

Menjelaskan proses desain aplikasi dan juga proses implementasi aplikasi dan juga implementasi metode K-Nearest Neighbor ke dalam sebuah sistem.

BAB 6 Pengujian dan Analisis

Pada bab ini akan menjelaskan pengujian sebuah sistem yang telah dibuat beserta analisis dari keseluruhan.

BAB 7 Kesimpulan

Menarik kesimpulan yang merupakan hasil dari pengujian/ pengolahan data dan analisis yang dilakukan.