

BAB I

PENDAHULUAN

1.1 Latar Belakang Penelitian

Media sosial memegang peranan yang sangat penting sebagai media komunikasi yang efektif dan efisien. Setiap orang bisa secara terbuka membagikan informasi dalam bentuk apapun seperti tulisan, gambar, video, maupun suara dalam waktu yang cepat dan tidak terbatas (Putri dkk, 2016). Kemudahan yang ditawarkan media sosial membuat jumlah penggunaanya terus bertambah setiap waktu. Menurut data dari (We Are Social, 2021) pengguna aktif media sosial di Indonesia per Januari 2021 mencapai 170 juta pengguna, dengan pertumbuhannya sebesar 6,3% dari tahun 2020.

Salah satu media sosial yang banyak digunakan adalah Twitter. Media sosial dengan fitur utama berbagi *tweet* berupa teks ini memiliki 14,05 juta pengguna aktif di Indonesia (We Are Social, 2021). Konsep dari Twitter adalah berbagi informasi berupa pesan singkat secara padat dan *real time* kepada pembacanya. Salah satu informasi yang sering dibagikan oleh pengguna Twitter adalah informasi dari kejadian apa yang sedang terjadi pada saat itu. Dengan menulis tweet mengenai kejadian yang sedang terjadi, informasi tersebut akan bermanfaat bagi pembacanya (Puspitadewi dkk, 2016). Selain bermanfaat bagi pembacanya, Twitter juga bisa dimanfaatkan bagi penulis tweet untuk meminta informasi kepada pengguna lainnya. Contohnya beberapa pengguna memanfaatkan Twitter untuk menulis tweet terkait kehilangan barang yang dialaminya, dengan harapan barang yang hilang bisa ditemukan oleh pembacanya. Sebaliknya, pengguna juga bisa menulis tweet terkait penemuan barang berharga dengan harapan pemiliknya mengetahui bahwa barangnya telah ditemukan.

Untuk membagikan informasi kehilangan perlu dibagikan kepada pengguna yang berdomisili di mana barang tersebut hilang untuk meningkatkan kemungkinan barang yang hilang bisa ditemukan. Cara agar bisa terhubung dengan para pengguna tersebut adalah dengan meminta bantuan akun media informasi yang berfokus pada penyampaian berita di daerah tertentu. Pengguna bisa menyebut

username akun media informasi tersebut pada tweet yang ditulis agar bisa dibagikan ulang kepada para pengikutnya.

Terdapat banyak akun yang dikhususkan sebagai media pemberi informasi di daerah tertentu. Salah satunya adalah akun @prfmnews yang dikelola oleh radio berita PRFM milik perusahaan media Pikiran Rakyat Group. Akun ini memberikan informasi berita dan kejadian yang terjadi di daerah kota Bandung dan sekitarnya. Sumber informasinya berasal dari pengikut atau pengguna yang melaporkan kejadian yang terjadi di sekitarnya, seperti lalu lintas, cuaca, bencana alam, kecelakaan, termasuk informasi berita kehilangan. Jumlah pengikutnya per Januari 2022 adalah sebanyak 275.000 dan telah menulis tweet sebanyak 704.000 sejak Juni 2010.

Tweet kehilangan yang dibagikan oleh @prfmnews akan membantu pengguna menemukan informasi mengenai barangnya yang hilang di sekitar kota Bandung karena memiliki banyak pengikut yang berdomisili di kota Bandung dan sekitarnya. Tweet yang dibagikan cukup beragam, seperti kehilangan dompet, dokumen, kendaraan, sampai kehilangan orang. Informasi tersebut bisa diklasifikasikan berdasarkan objek kehilangannya. Model klasifikasi yang telah dibuat bisa digunakan untuk pembuatan sistem informasi kehilangan.

Tabel 1.1 Contoh *tweets* kehilangan

| Tweet | Jenis Kehilangan |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|
| via @ipa_iiiipa Mohon bantuannya, telah hilang sebuah dompet. Isinya ada KTP dan SIM atas nama Rosipa, STNK, ATM, kartu berobat, sama kartu lainnya. Ada uangnya tapi ga banyak. Kemungkinan jatuh disekitar paskal, kebon jati, pasar baru, jln stasiun, perintis, atau wastukencana. | Barang dan Dokumen |
| RT @SalamMulki: Maaf bantu retweet teman. Telah hilang motor beat warna hitam daerah Sukajadi Bandung berplat nomor D 2608 ACB, barangkali ada yang melihat atau menemukan tiasa DM ka @SalamMulki. Hatur Nuhun. | Kendaraan |
| INFO ANAK HILANG. Nama : Kevin. Umur 14th. Baru tiba di Bandung 2 hari lalu dari Sumatera Selatan (Kab Muara Enim) tadi pukul 14.30 kabur dari rumah, alamat Cilengkrang 2 jln panorama indah no 8, rt6/rw14 Kel Palasari, kec cibiru. | Orang |
| DPRD Minta Dishub Gencarkan Penggunaan Transportasi Umum Biar Kemacetan di Kota Bandung Hilang | Bukan Informasi Kehilangan |

Pada tabel 1.1 dijelaskan contoh dari *tweets* yang diperoleh dari akun @prfmnews dengan kata kunci ‘hilang’. *Tweet* pertama berisi informasi kehilangan barang dan dokumen, yaitu dompet beserta surat-surat penting lainnya. *Tweet* kedua berisi informasi kehilangan kendaraan berupa motor. *Tweet* ketiga merupakan informasi kehilangan orang yang diduga kabur dari rumah. *Tweet* keempat bukan merupakan *tweet* kehilangan, karena tidak memuat informasi mengenai hilangnya suatu barang atau seseorang. Masing-masing *tweet* tersebut merepresentasikan kelas yang bisa digunakan untuk pelabelan *tweet* kehilangan lainnya sehingga klasifikasi bisa dilakukan.

Namun, klasifikasi dari data teks memiliki masalah yang berkaitan dengan banyaknya atribut yang digunakan karena data teks memiliki kosakata yang sangat banyak. Jika semua atribut digunakan dalam klasifikasi, hal ini akan mengurangi performa dari *classifier* (Chandani dkk, 2015). Oleh karena itu diperlukan

pemilihan atribut yang memiliki pengaruh besar dan merupakan ciri dari kelas yang terkait. Pemilihan atribut atau *feature selection* bisa meningkatkan efisiensi pengklasifikasian karena parameter yang perlu diestimasi dan perhitungan yang dilakukan oleh *classifier* menjadi lebih sedikit. Selain itu, *feature selection* juga dapat meningkatkan akurasi jika atribut yang dipilih sudah tepat (Schneider, 2005).

Metode *feature selection* yang akan digunakan dalam penelitian ini adalah Information Gain (IG), Chi-Square (X²), dan Document Frequency (DF). Penelitian yang dilakukan oleh Yang dan Pedersen (1997) menemukan bahwa Information Gain dan Chi-Square merupakan metode *feature selection* paling efektif dalam mengeliminasi fitur tanpa kehilangan akurasi dari klasifikasi dalam eksperimen yang mereka lakukan dengan menghapus hingga 98% dari fitur. Kemudian metode Document Frequency juga memiliki performa yang sebanding saat fitur dihapus sebanyak 90%. Selain itu, Document Frequency juga digunakan untuk menghapus fitur yang bersifat *noise* karena menghitung frekuensi kemunculan fitur.

Metode tersebut akan digunakan sebagai perhitungan untuk mendapatkan fitur yang paling berpengaruh terhadap dataset *tweet* kehilangan. Sedangkan metode klasifikasi yang akan digunakan adalah Naïve Bayes. Naïve Bayes menggunakan model probabilistik sederhana yang memungkinkan untuk menyimpulkan kelas dari dokumen yang tidak diketahui menggunakan aturan Bayes. Karena tekniknya yang sederhana, metode ini banyak digunakan dalam klasifikasi teks (Schneider, 2005). Kemudian pembobotan pada penelitian ini akan menggunakan TFIDF. Hal ini dilakukan untuk menguji metode Naïve Bayes yang menggunakan penghitungan probabalistik berdasarkan frekuensi kemunculan suatu kata, apakah jika dilakukan penghitungan dengan bobot nilai riil yang dimiliki TFIDF akan meningkatkan performa akurasi atau tidak.

Sebelum melakukan klasifikasi, data akan di-praproses terlebih dahulu untuk menyederhanakan kosakata yang akan digunakan sebagai atribut. Tahap yang pertama adalah dengan mengubah semua huruf menjadi huruf kecil, karena mesin

tidak bisa menggeneralisir huruf besar dan kecil. Kemudian karakter yang berbentuk simbol dan angka akan dihapus karena yang diperlukan dalam klasifikasi hanya karakter huruf yang membentuk kata. Setelah itu dilakukan proses stemming untuk menggeneralisir kata dengan menghilangkan setiap imbuhan yang digunakan. Contohnya kata “melaporkan” dan “dilaporkan” akan menjadi sebuah kata yaitu “lapor”. Kemudian tahapan selanjutnya adalah menghapus stopwords, yaitu Setelah itu dilakukan feature selection dengan empat metode yang sudah disebutkan. Data yang sudah melalui proses feature selection masing-masing akan diklasifikasikan. Kemudian hasilnya akan dibandingkan performa dari masing-masing metode

1.2 Rumusan Masalah Penelitian

Berdasarkan latar belakang masalah, maka dapat dirumuskan masalah yang akan dibahas dalam penelitian ini adalah sebagai berikut.

1. Bagaimana implementasi teknik *feature selection* menggunakan metode Information Gain, Chi-Square, dan Document Frequency untuk mengklasifikasikan informasi kehilangan pada data Twitter?
2. Bagaimana implementasi pembobotan fitur menggunakan TFIDF untuk mengklasifikasikan informasi kehilangan pada data Twitter?
3. Bagaimana perbandingan performa dari klasifikasi berita kehilangan pada data Twitter menggunakan Naïve Bayes dengan penerapan *feature selection* dan pembobotan TFIDF?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan dilakukannya penelitian ini adalah sebagai berikut.

1. Mengimplementasikan teknik *feature selection* menggunakan metode Information Gain, Chi-Square, dan Document Frequency untuk mengklasifikasikan informasi kehilangan pada data Twitter.
2. Mengimplementasikan pembobotan fitur menggunakan TFIDF untuk mengklasifikasikan informasi kehilangan pada data Twitter.

3. Membandingkan performa dari klasifikasi informasi kehilangan pada data Twitter menggunakan Naïve Bayes dengan penerapan *feature selection* dan pembobotan TFIDF.

1.4 Manfaat Penelitian

Dengan dilakukannya penelitian ini, diharapkan dapat memberikan beberapa manfaat, diantaranya:

1. Dapat memberikan pengetahuan mengenai dampak penggunaan teknik *feature selection* untuk mengklasifikasikan data teks.
2. Dapat memberikan pengetahuan mengenai dampak penggunaan TFIDF untuk pembobotan fitur.
3. Dapat memberikan kemudahan bagi pengguna untuk mencari informasi mengenai kehilangan atau penemuan barang.
4. Dapat memberikan masukan untuk pembuatan sistem informasi kehilangan.

1.5 Batasan Masalah

Batasan masalah sangat diperlukan agar pembahasan tidak terlalu luas, karena studi kasus yang diteliti begitu luas maka dalam hal penelitian menggunakan batas-batas sebagai berikut.

1. Data yang digunakan diambil dari *tweet* yang di-*posting* oleh akun @prfmnews dan *tweet* pengguna yang menyebutkan akun @prfmnews.
2. Data yang digunakan hanya *tweet* yang memiliki kata kunci “kehilangan”, “hilang”, “ditemukan”, “menemukan”, “pencurian”, dan “dicuri”.
3. Data yang digunakan diambil dari rentang waktu Januari 2019 sampai Desember 2021.

1.6 Struktur Organisasi Penelitian

Pada bagian sistematika penulisan ini akan diuraikan mengenai penjelasan tiap bab.

BAB I PENDAHULUAN

Pada bab ini dijelaskan latar belakang dilakukannya penelitian “Implementasi *Feature Selection* dan Pembobotan TFIDF Pada Klasifikasi *Tweets* Informasi Kehilangan Dengan Naïve Bayes” berdasarkan permasalahan yang terjadi, yaitu informasi kehilangan yang tidak terklasifikasi. Dari permasalahan tersebut kemudian memunculkan gagasan untuk membuat program dengan mengimplementasikan metode-metode yang relevan. Selanjutnya permasalahan yang akan dipecahkan diuraikan dalam rumusan masalah penelitian. Berdasarkan rumusan masalah tersebut, diuraikan juga tujuan yang ingin dicapai dari penelitian ini. Seiring dengan tercapainya tujuan penelitian, manfaat yang dapat diperoleh dari penelitian ini juga dijabarkan. Kemudian dalam melakukan penelitian akan ditentukan batasan-batasan agar penelitian terfokus ke tujuan yang ingin dicapai.

BAB II KAJIAN PUSTAKA

Pada bab ini dijelaskan studi literatur dari teori-teori sebagai basis ilmu dalam penelitian ini. Gambaran studi literatur dijelaskan menggunakan peta literatur berdasarkan teori apa saja yang akan ditulis dalam bab ini. Studi literatur pertama dilakukan untuk menjelaskan media sosial Twitter mengenai sejarah, profil, dan konsepnya. Kemudian pada studi literatur selanjutnya berkaitan dengan teknik pemrosesan teks di mana data berbentuk teks perlu dibersihkan terlebih dahulu sebelum diproses. Setelah itu dijelaskan metode *feature selection* dari Information Gain, Chi-Square, dan Document Frequency untuk menghitung nilai dari masing-masing fitur. Selanjutnya dilakukan kajian pada Pembobotan TFIDF mengenai teknik untuk memberikan bobot fitur di setiap dokumen. Kemudian dijelaskan juga mengenai metode klasifikasi yang digunakan yaitu Naïve Bayes. Pada subbab tersebut dijelaskan pengertian, rumus, dan cara penghitungannya. Yang terakhir dipaparkan penelitian-penelitian yang terkait sebagai referensi dan gambaran dari penelitian yang akan dilakukan

BAB III METODOLOGI PENELITIAN

Pada bab ini dijelaskan tahapan-tahapan yang akan dilakukan pada penelitian ini. Desain penelitian digambarkan untuk menjelaskan alur dari penelitian, dimulai dari perumusan masalah sampai analisis dan kesimpulan. Untuk mengimplementasikan alur pada desain penelitian, perlu dijabarkan juga metode penelitian yang digunakan, yaitu metode pengumpulan data dan metode pengembangan perangkat lunak. Kemudian untuk menjalankannya perlu diuraikan juga perangkat dan data penelitian yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi penjelasan bagaimana implementasi dalam membangun program berdasarkan tahapan dalam metode pengembangan perangkat lunak yang digunakan. Tahapan dimulai dengan tahap analisis untuk menjelaskan analisis yang dilakukan sebelum program dibangun. Tahapan ini terbagi menjadi tiga, yaitu analisis umum, analisis data, dan analisis kebutuhan. Kemudian dilanjutkan dengan tahap perancangan, yaitu proses menggambarkan alur dari program yang dibangun. Tahapan ini terbagi menjadi tiga tahap, yaitu alur pengumpulan data, alur *training* data, dan alur *testing* data. Selanjutnya alur yang telah digambarkan dalam perancangan akan diimplementasikan pada tahap implementasi. Tahapan ini menjelaskan proses pembangunan program, dimulai dari persiapan data hingga melakukan evaluasi. Kemudian dijelaskan pengujian dari program yang dibangun pada tahap *testing*. Setelah program berhasil dibangun, dilakukan eksperimen untuk mendapatkan hasil berdasarkan tujuan dari penelitian. Eksperimen yang telah dilakukan kemudian dilakukan analisis untuk menjelaskan hasil dari eksperimen.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari penelitian yang telah dilakukan sesuai dengan tujuan dari penelitian. Kesimpulan juga ditulis berdasarkan analisis dari hasil pengujian yang dilakukan pada bab hasil dan pembahasan. Lalu disertakan juga saran yang bisa dilakukan untuk penelitian selanjutnya.