BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Pertumbuhan pengguna internet terus meningkat sesuai dengan berjalannya waktu. Terhitung mulai dari tahun 1993 hingga 2016 pengguna internet mencapai 3,4 miliar dengan tingkat penetrasi 46%[1]. Tingkat penetrasi tersebut menunjukkan bahwa satu dari dua orang di dunia memiliki akses internet. Pertumbuhan pengguna internet didorong untuk menjawab berbagai macam kebutuhan mulai dari hiburan, pendidikan, bisnis, pelayanan, hingga transportasi. Pertumbuhan pengguna internet memberikan dampak pada pertumbuhan data secara signifikan. Tercatat pada tahun 2016 rata-rata dalam setiap hari 200 ribu blog terbit di wordpress, 3 miliar pencarian di google, 500 juta tweet terkirim, 40 miliar foto dibagikan di instagram, 350 juta foto diunggah di facebook[2]. Data tersebut akan menjadi sampah jika hanya disimpan tanpa dianalisis lebih lanjut. Salah satu teknik analisis yang bisa dilakukan adalah analisis sentimen. Teknik tersebut dapat menganalisis pendapat, sentimen, evaluasi, penilaian, sikap, dan emosi masyarakat terhadap entitas seperti produk, layanan, organisasi, individu, isu, peristiwa, topik, dan atribut. Selain analisis sentimen, teknik tersebut dikenal juga dengan istilah berbeda, misalnya, *opinion mining, opinion extraction, sentimen mining, subjectivity analysis, affect analysis, emotion analysis, review mining*, dan lain-lain. Dari beragam istilah tersebut, kalangan akademisi dan industri lebih sering menggunakan istilah *sentiment analysis*[3].

Sejarah analisis sentimen pertama kali muncul pada sebuah jurnal karya Das, Chen, dan Tong [4] pada tahun 2001. Bahasan yang diangkat yaitu menganalisis sentimen pasar. Sejumlah karya lainnya menyebutkan bahwa analisis sentimen berfokus pada aplikasi spesifik yang mengklasifikasikan mengenai sifat yang berlawanan yakni antara positif dan negatif. Pada tahun 2002, Bo Pang [5] melakukan analisis sentimen menggunakan data ulasan film dengan memanfaatkan metode *machine learning*. Kemunculan metode *machine learning* didorong oleh tersedianya dataset

dari ledakan data di internet dan kebutuhan industri hingga kini penelitian analisis sentimen terus berkembang.

Kini analisis sentimen telah banyak digunakan di berbagai industri, contohnya adalah disektor *e-commerce*. Analisis sentimen digunakan untuk mengetahui seberapa jauh kepuasan pembeli dari produk dan layanan yang telah diberikan. Sumber data yang digunakan untuk mengetahui tingkat kepuasan pembeli dapat berupa opini. Hasil dari analisis sentimen tersebut dapat dijadikan oleh perusahaan sebagai alat evaluasi untuk meningkatkan pelayanan. Selain *e-commerce*, dunia politikpun kini sudah mulai memanfaatkan analisis sentimen sebagai alternatif alat ukur popularitas dan elektabilitas politikus atau tokoh lainnya dan juga digunakan untuk melihat reaksi publik terhadap isu yang ada.

Entitas yang akan diukur sentimen publik menggunakan analisis sentimen dapat menggunakan dataset yang dimiliki oleh entitas tersebut (data internal). Selain data internal, kini bisa memanfaatkan data eksternal yang didapat dengan menggunakan teknik *scrapping* yaitu mengambil data dari suatu situs atau menggunakan *application programming interface* (API) yang disediakan media sosial seperti facebook, instagram, atau twitter. Ketiga media sosial tersebut termasuk ke dalam media sosial yang memiliki pengguna aktif paling banyak [6]. Di Indonesia, facebook memiliki sekitar 106 juta pengguna atau sekitar 6% dari seluruh pengguna facebook berasal dari Indonesia [7]. Dengan jumlah pengguna sebesar itu, opini dengan topik bisnis hingga politik dapat ditemukan di facebook.

Dalam perkembangannya, algoritma yang digunakan untuk melakukan analisis sentimen terus dikembangkan untuk menaikkan tingkat ketepatan dalam melakukan klasifikasi (akurasi) sentimen suatu opini. Di beberapa penelitian, algoritma *Support Vector Machine* (SVM) dinilai memiliki tingkat akurasi yang paling tinggi di antara algoritma yang lainnya. Tim O'Keefe dan Irena Koprisnka [8] pada penelitiannya tahun 2006 menunjukkan SVM lebih unggul dibanding *Naïve Bayes Classifier* (NBC). Dalam penelitian yang dilakukan pada tahun 2009 oleh Fatimah Wulandini dan Anto Satriyo Nugroho [9] menunjukkan bahwa SVM memiliki

tingkat akurasi lebih tinggi dibanding dengan algoritma NBC, C4.5 dan *K-Nearest Neighbor*.

Sebelum algoritma SVM dapat melakukan klasifikasi terhadap suatu dokumen berdasarkan sentimennya, ada tahapan yang lebih awal dilakukan yaitu *term weighting* (pembobotan) yang merubah kata-kata menjadi angka. Transformasi pembobotan dari kata-kata menjadi angka didasari pada kemunculannya. Ada yang memperhitungkan kemunculannya berdasarkan label dari dokumen (*supervised term weighting*), ada yang tanpa memperhitungkan kemunculannya berdasarkan label dari dokumen (*unsupervised term weighting)*. Pemilihan dua teknik pembobotan tersebut dapat mempengaruhi tingkat akurasi yang akan dihasilkan oleh algoritma SVM. Penelitian mengenai pemilihan teknik pembobotan untuk SVM telah dilakukan oleh Man Lan dan Chew Lim Tan pada tahun 2007[10]. Penelitian ini menunjukkan bahwa teknik pembobotan *Term Frequency Odd Ratio* (TF-OR) memiliki tingkat tertinggi dibanding teknik pembobotan lainnya. Hasil tersebut dicapai dengan catatan komposisi data latih positif dan negatif yang digunakan seimbang, sebagai contoh data latih memiliki 50% dokumen berlabel negatif dan 50% dokumen berlabel positif.

Ketika TF-OR menghadapi dataset yang tidak seimbang, akurasi yang didapat menjadi lebih rendah dibanding dengan teknik pembobotan *Term Weighting – Inverse Document Frequency* (TF-IDF) yang tidak mempertimbangkan kemunculan suatu kata di dokumen positif atau negatif [11]. Maka penelitian ini bermaksud untuk membuat suatu model pembobotan dengan memanfaatkan TF-OR dan TF-IDF sehingga membuat akurasi algoritma SVM menjadi relatif stabil ketika menggunakan data latih yang seimbang atau tidak seimbang.

1.2 Rumusan Masalah

Berdasarkan latar belakang, berikut adalah rumusan masalah yang diperoleh dalam penelitian ini:

1. Diperlukan teknik pembobotan baru yang dapat menghasilkan akurasi SVM tinggi saat data latih seimbang atau tidak seimbang.

2. Berapa penentuan rasio ideal data latih dan data uji yang ideal untuk menghasilkan analisis sentimen yang baik.

1.3 Premis Penelitian

Berdasarkan rumusan masalah, berikut adalah premis-premis yang diperoleh dalam penelitian ini:

1. Teknik pembobotan TF-OR akan membantu algoritma SVM menghasilkan tingkat akurasi yang lebih tinggi ketika komposisi data latih seimbang.
2. Teknik pembobotan TF-IDF akan membantu algoritma SVM menghasilkan tingkat akurasi yang lebih tinggi ketika komposisi data latih tidak seimbang.

1.4 Hipotesis Penelitian

Penggabungan teknik pembobotan TF-OR dan TF-IDF dapat meningkatkan akurasi algoritma SVM dengan komposisi data latih yang seimbang atau dengan komposisi data latih yang tidak seimbang.

1.5 Batasan Masalah

Pembahasan dalam penelitian ini dibatasi agar tidak menyimpang dari tujuan yang ingin dicapai, adapun batasan masalahnya adalah:

1. Dataset yang digunakan menggunakan ulasan film dari imdb [12].
2. Menggunakan dua kelas sentimen yaitu positif dan negatif.
3. Data uji yang digunakan berasal dari dataset yang dibagi menjadi data latih dan data uji.
4. Pengukuran akurasi menggunakan *confusion matrix* [13].

1.6 Manfaat Penelitian

Berdasarkan maksud dan tujuan dari penelitian ini maka akan didapat manfaat berupa diketahuinya model pembobotan TF-IDFOR yang memiliki pendekatan pendeteksian kondisi (seimbang atau tidak seimbang) data latih untuk menentukan pembobotan yang tepat untuk meningkatkan tingkat akurasi dari algoritma SVM. Selain itu akan diketahui pula kondisi data latih yang lebih optimal untuk TF-OR dibanding dengan TF-IDF begitu juga sebaliknya dan juga mengetahui rasio antara

data latih dan data uji yang ideal untuk melakukan inisialisasi analisis sentimen berikut dengan kuantitas idealnya.

1.7 Sistematika Penulisan

Sistematika penulisan penelitian ini disusun untuk memberikan gambaran umum mengenai penelitian yang dikerjakan. Berikut adalah sistematika penulisan yang digunakan:

BAB I PENDAHULUAN

Bab ini membahas latar belakang permasalahan, merumuskan inti permasalahan, mencari solusi dari permasalahan tersebut, merumuskan masalah tersebut, menentukan maksud dan tujuan, membatasi masalah penelitian, membahas manfaat yang diperoleh dari hasil penelitian, dan menentukan sistem penulisan yang digunakan,

BAB II LANDASAN TEORI

Bab ini membahas berbagai konsep dasar dan teori-teori yang berkaitan dengan analisis sentimen. Mulai dari algoritma yang digunakan, teknik pembobotan, hingga teknik evaluasi dari hasil analisis sentimen.

BAB III METODE PENELITIAN

Bab ini membahas desain penggabungan dua teknik pembobotan TF-IDF dan TF-OR, data penelitian yang digunakan, dan alur pengujian.

BAB IV HASIL PENELITIAN

Bab ini membahas hasil desain yang telah dibuat. Dalam bab ini akan diuraikan hasil analisis sentimen menggunakan TF-IDF dengan SVM dan TF-OR dengan SVM berdasarkan alur pengujian yang telah dibuat.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari semua hal yang dibahas pada bab sebelumnya dan saran untuk tahap penelitian selanjutnya.