BAB I
PENDAHULUAN

1.1 Latar Belakang Penelitian

Teknologi informasi dan komunikasi telah memberikan banyak kemudahan dalam mengelola dan penyebaran informasi. Teknologi informasi tersebut banyak dimanfaatkan pada penggunaan halaman *website*. Berdasarkan *Netcraft Web Server Survey* jumlah halaman website yang aktif pada Februari 2018 mencapai 838 milyar situs web. Peningkatan jumlah halaman *website* tersebut berdampak pada peningkatan jumlah dokumen dari berbagai penyedia sumber informasi.

Peningkatan jumlah dokumen dari berbagai sumber informasi, salah satunya berdampak pada halaman penyedia jurnal internasional. Dimana penyedia informasi tersebut memanfaatkan halaman *website* dalam sarana penyebaran jurnal sebagai bahan penunjang penelitian dan dapat dengan mudah di akses dari berbagai negara. Bertambahnya jumlah dokumen jurnal akan membuat masyarakat umum mengalami kesulitan dalam menemukan dokumen jurnal sesuai dengan keinginan. Oleh karena itu diperlukan teknik pengolahan teks yang mengorganisasikan dokumen teks dalam jumlah besar sesuai dengan kategorinya, sehingga informasi yang tersedia dapat terorganisasi dengan baik dan mudah di akses sesuai dengan kebutuhan pengguna.

Pemecahan masalah dalam pengkategorian dokumen teks dapat diselesaikan dengan menggunakan *text mining* (Wiyonto, 2016). *Text mining* merupakan variasi dari *data mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual dalam jumlah yang besar (Fithriasari, 2016). Algoritma dalam *text mining* dapat mengenali data semi terstruktur dari dokumen. *Text mining* dapat menyelesaikan permasalahan klasifikasi dengan menganalisis dan menggali informasi dalam teks yang diilakukan secara otomatis. Selain itu *text mining* dapat menyelesaikan permasalahan lain seperti *clustering*, *information extraction* dan *information retrival* (Hamim, 2016).

Pemanfaatan *text mining* untuk pengklasifikasian data akan membuat pengelolaan informasi menjadi efektif (Somantri, 2016). Algoritma atau metode untuk klasifikasi dokumen teks telah banyak berkembang baik berupa algoritma

klasifikasi *supervised learning* maupun *unsupervised learning*. Perbedaan dari kedua algoritma tersebut adalah algoritma *supervised learning* menggunakan data latih sebagai metode pembelajarannya atau dalam pembuatan model *classifier*, sedangkan untuk algoritma *unsupervised learning* tidak menggunakan data latih sebagai tahapan pembuatan model *classifier*. Sehingga dengan algoritma klasifikasi *supervised learning* data-data sebelumnya atau data latih akan memiliki variabel target yang akan diklasifikasikan (Chandara, 2017).

Penelitian pengklasifikasin dokumen teks sebelumnya telah dilakukan oleh Wahid pada tahun 2016 dalam melakukan pengklasifikasian dokumen twitter. Penelitian tersebut menggunakan algoritma klasifikasi *unsupervised learning* yaitu SentiStrength dengan menggunakan *hybrid* TF-IDF dan *cosine similarity*. Dari penelitian tersebut terlihat bahwa algoritma SentiStrength hanya menggunakan bebebrapa fitur dalam proses klasifikasi, sehingga performansi yang dihasilkan masih belum maksimal. Penelitian tersebut menggunakan 30% dan 50% data uji dan menghasilkan rata-rata presisi adalah 60%, *recall* 65% dan *f1-measure* 62%.

Wahid (2016) mengusulkan untuk dilakukan pembangunan sistem klasifikasi dengan teknik *supervised learning*. Hal itu karena teknik tersebut dalam pembangunan model klasifikasi dilakukan berdasarkan data latih pembelajaran yang dapat dengan mudah membuat kontrol *training sample* terhadap *information classes.* Sehingga hasil klasifikasi menjadi lebih baik dari teknik *unsupervised learning*. Beberapa algoritma *supervised learning* diantaranya *K-Nearest Neighbour* (KNN), *Naïve Bayes Classifier*, *Decision Tree*, *Support Vector Machines* dan lain-lain.

Algoritma *K-Nearest Neighbour* (KNN) dapat melakukan klasifikasi dengan cepat berdasarkan jarak terdekat diantara objek data (Liantoni *et al*, 2015). Pengklasifikasian dokumen menggunakan algoritma *K-Nearest Neighbour* (KNN) bergantung dari nilai *k* (jumlah tetangga) yang digunakan. Untuk nilai *k* (jumlah tetangga) yang digunakan tidak memiliki batasan. Hal tersebut untuk menguji model classifier dalam menemukan hasil performansi yang maksimal. *K-Nearest Neighbour* (KNN) merupakan metode pengklasifikasian yang tangguh terhadap data training yang memiliki banyak *noise* dan keefektifan apabila data *training* besar (Ilyas, 2009). Dari berbagai kasus beberapa penelitian terdahulu untuk *text*

classification metode *K-Nearest Neighbour* (KNN) menunjukan performasi yang lebih baik dari algoritma lain. Seperti penelitian yang dilakukan oleh Muhammad Fakhrurrifqi *et al* (2013) yang membandingkan algoritma klasifikasi *K-Nearest Neighbour* (KNN) dengan C4.45 dan LVQ, hasil menunjukan bahwa algoritma *K-Nearest Neighbour* (KNN) memberikan ketepatan yang lebih baik dan akurasi yang lebih tinggi.

Berdasarkan pada hasil penelitian sebelumnya maka akan digunakan algoritma klasifikasi *K-Nearest Neighbour* (KNN) untuk mengklasifikasikan jenis jurnal internasional ilmu komputer. *K-Nearest Neighbour* (KNN) dapat diimplementasikan dengan mudah, dapat memisahkan lebih dari dua kelas atau *multiclass* dan metode ini paling banyak dipakai untuk klasifikasi data teruma data teks. Pengklasifikasian dokumen teks dengan algoritma klasifikasi *supervised learning* menggunakan pemodelan *classifier* dari data pembelajaran.

Pembentukan model *classifier* meliputi pembagian data latih (data *training*) dan data uji (data *testing*). Data *training* berfungsi sebagai pembentukan model klasifikasi sedangkan data *testing* berfungsi untuk menguji model klasifikasi yang telah di buat. Terdapat beberapa metode pembagian data diantaranya yang popular digunakan yaitu metode pembagian data dengan persentase atau *holdout method* dan metode 10 *fold cross validation.* Untuk mengetahui model *classifier* terbaik, penelitian ini menggunakan kedua metode tersebut dalam pembagian data untuk pembentukan *classifier.*

Maka dalam penelitian ini akan dilakukan klasifikasi dokumen teks, yang terlebih dahulu dokumen melalui beberapa tahapan diantaranya *preprocessing text* dan *feature selection* (Wiyonto, 2016). *Preprocessing text* dilakukan untuk melakukan pembersihan data sebelum diklasifikasikan yang terdiri dari 3 tahap yaitu pembersihan karakter yang tidak dibutuhkan, mengubah bentuk kalimat atau paragraf menjadi bentuk *term* atau kata dan menghilangkan *term* yang tidak berpengaruh besar pada proses klasifikasi. Kemudian *feature selection* dilakukan untuk menganalisis *term* pada setiap dokumen yaitu dengan melakukan pembobotan TF-IDF (Hamim, 2016). *Feature selection* diperlukan karena memiliki kemampuan untuk memilih atribut yang penting dan dapat membuang atribut yang tidak terkait atau duplikasi (Wibisono, 2012).

Dalam penelitian ini akan dilakukan proses sebuah sistem klasifikasi dokumen teks berbahasa Inggris dengan tahapan *preprocessing text* dan *feature selection.* Algoritma klasifikasi yang digunakan adalah *supervised learning* yaitu *K-Nearest Neighbor* (KNN) dengan *cosine silimarity.* Untuk mengetahui model *classifier* terbaik dari sistem klasifikasi yang dilakukan, peneliti menggunakan 2 metode pembagian data yang berbeda yaitu dengan *holdout method* dan 10 *fold cross validation*.

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian yang telah diuraikan pada Sub Bab 1.1, rumusan masalah sebagai berikut :

1. Bagaimana merancang sistem klasifikasi dokumen abstrak jurnal internasional ilmu komputer menggunakan algoritma klasifikasi *K-Nearest Neighbor* (KNN) ?
2. Bagaimana perbedaan hasil implementasi sistem klasifikasi pada dokumen abstrak jurnal internasional ilmu komputer dengan metode pembagian data *holdout method* dan 10 *fold cross validation* ?
3. Bagaimana perbandingan hasil implementasi sistem klasifikasi dokumen abstrak jurnal internasional ilmu komputer dengan penelitian sebelumnya ?

1.3 Tujuan Penelitian

Setelah diketahui rumusan masalahnya, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Merancang sistem klasifikasi pada dokumen jurnal internasional ilmu komputer menggunakan algoritma klasifikasi *K-Nearest Neighbor* (KNN).
2. Mengetahui perbedaan hasil implementasi sistem klasifikasi abstrak jurnal internasional ilmu komputer antara dua metode pembagian data yang berbeda yaitu *holdout method* dan 10 *fold cross validation.*
3. Mengetahui pemodelan *classifier* terbaik dari skenario pembagian data antara *holdout method* dan 10 *fold cross validation.*
4. Mengetahui perbandingan hasil eksperimen dari penelitian yang dilakukan dengan penelitian sebelumnya.

1.4 Manfaat Penelitian

Adapun manfaat penelitian ini, diantaranya sebagai berikut :

1. Mempermudah dalam mengelompokan dokumen jurnal internasional ilmu komputer secara otomatis berdasarkan kategori tertentu dalam jumlah data yang besar.
2. Menambah alur sistem pengklasifikasian dokumen berbahasa Inggris dan membantu dalam mengorganisasikan dokumen secara cepat, efisien dan memiliki kinerja yang baik.
3. Mengetahui kelebihan dan kekurangan metode algoritma klasifikasi *supervised learning* yaitu *K-Nearest Neighbor* (KNN) dalam klasifikasi data teks berbahasa Inggris.

1.5 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Sistem klasifikasi jurnal dikembangkan hanya untuk jurnal berbahasa Inggris dengan data abstrak sebagai acuan pengklasifikasian.
2. Pelabelan data diberikan berdasarkan label yang tertera pada situs ScienceDirect diantaranya yaitu kelas *Computer and Education, Computer and Security* dan *Computer in Human Behavior.*

1.6 Sistematika Penulisan

Sistematika penulisan ini akan diuraikan mengenai penjelasan tiap bab.

BAB I PENDAHULUAN

Bab ini berisi bagaimana awal mula penelitian muncul dan pembahasannya mengenai konteks penelitian yang dilakukan, diawali dengan latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah dan sistematika penulisan.

BAB II KAJIAN PUSTAKA

Bab ini berisi berbagai kajian teori pendukung untuk melakukan penelitian. Kajian teori ini akan memudahkan dalam pemenuhan kebutuhan bahan penelitian. Teori yang dijelaskan dalam bab ini yaitu : *data mining*, *text mining* yang meliputi *text preprocessing* dan pembobotan TF-IDF, *splitting*

data dengan *holdout method* dan 10 *fold cross validation*, *text classification*, cosine similarity, K-Nearest Neighbor (KNN), jurnal penelitian ilmu komputer, penelitian terkait KNN *classification*, dan evaluasi hasil klasifikasi.

BAB III METODOLOGI PENELITIAN

Bab ini berisi penjelasan langkah-langkah penelitain yang akan dilakukan, dengan membuat desain penelitian sebagai gambaran dalam langkah-langkah analisis penelitian dan proses pengklasifikasian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil analisis pengklasifikasian yang terdiri dari pengumpulan data, perancangan *data preparation*, perancangan model *classifier*, implementasi sistem klasifikasi, perbandingan dengan penelitian sebelumnya.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil analisis yang telah dilakukan dan diikuti dengan saran pengembangan penelitian