BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan Teknologi Informasi saat ini sangat berpengaruh terhadap berbagai bidang dalam kehidupan sehari-hari manusia. Tak terkecuali dalam bidang penyebaran informasi yang kini semakin mudah. Berita yang dahulu banyak disebarkan melalui media cetak sekarang mulai memasuki atau bahkan berpindah pada media *online* untuk penyebarannya. Media *online* memiliki berberapa kelebihan, salah satunya adalah berita yang disebar melalui media *online* memiliki waktu distribusi yang lebih cepat dibandingkan dengan media cetak. Hal tersebut menyebabkan terjadinya membeludaknya informasi yang tersedia di internet.

Pada saat ini pengkategorian berita masih menggunakan tenaga manusia atau manual. Kategori yang banyak beserta waktu yang cepat akan menyulitkan editor untuk mengkategorikan berita, terutama berita yang tidak terlalu berbeda secara jelas. Beberapa kategori yang penggunaan bahasanya tidak berbeda terlalu jauh seperti nasional, internasional, pengetahuan, ekonomi, teknologi, kesehatan, dan properti mengharuskan seorang editor mengetahui isi berita yang akan diunggah secara keseluruhan untuk selanjutnya dimasukkan ke dalam kategori yang tepat. Akan lebih efesien apabila kategori berita dimasukkan secara otomatis dengan komputer menggunakan metode tertentu. (Ariadi & Fithriasari, 2015)

Sebelum berita masuk ke proses klasifikasi, sebelumnya harus dilakukan *text preprocessing* pada berita. Dalam proses ini terdi dari beberapa proses seperti

case folding yaitu proses untuk mengubah semua karakter pada teks menjadi huruf kecil, *tokenizing* untuk memecah kalimat menjadi kata per kata, *stemming* pada kata-kata yang tersisa pada dokumen teks untuk mendapatkan kata dasar, terakhir dilakukan proses *stopping* berdasarkan *stoplist* yang berisi *stopwords* yang telah ditentukan sebelumnya. (Ariadi & Fithriasari, 2015)

Selanjutnya harus dilakukan ekstraksi fitur pada berita. Dalam beberapa tahun terakhir, penggunaan *Term Frequency – Inverse Document Frequency* (*TF-IDF*) masih populer walaupun data dalam jumlah besar diekstraksi dari *TF-IDF*. *TF-IDF* adalah metode statistik numerik yang memungkinkan penentuan bobot untuk setiap istilah (atau kata) dalam setiap dokumen. Metode ini menentukan bobot, ukuran yang mengevaluasi pentingnya istilah (atau kata-kata) dalam pengumpulan dokumen. (Wibowo, Kartika, & Wardhana, 2018)

Pengkategorian berita secara otomatis dapat dilakukan dengan cara penerapan metode klasifikasi. Terdapat banyak metode klasifikasi yang dapat dipakai untuk melakukan pengkategorian berita diantaranya adalah *Naïve Bayes Classifier* (*NBC*), *Support Vector Machine* (*SVM*), *K-Nearest Neighbor* (*KNN), Neural Network Classifier*, serta *Probabilistic Latent Semantic Analysis (PLSA). NBC* dan *SVM* adalah metode yang paling banyak digunakan untuk menyelesaikan masalah klasifikasi dokumen dan keduanya memberikan hasil yang cukup menjanjikan. Perbandingan antara kedua metode *NBC* dan *SVM* didapatkan hasil *SVM* dengan kernel *Radial Basis Function* (*RBF)* dan linier lebih baik dibandingkan dengan *NBC*, selain itu waktu proses program saat menggunakan *SVM* juga jauh lebih cepat untuk mendapatkan hasil daripada *NBC*. (Ariadi & Fithriasari, 2015)

NBC memiliki banyak celah dalam efektifitas, misalnya dalam kasus *spam filtering* dengan memasukkan kata-kata asing sehingga perangkat lunak tidak dapat melakukan pengecekan, atau dengan memasukkan banyak kata yang sebenarnya sering digunakan oleh surat elektronik bukan *spam* supaya dapat memperkecil nilai probabilitas kata-kata *spam*. (Natalius, 2010) Sementara *SVM* memiliki kemampuan untuk mengklasifikasikan suatu *pattern*, yang tidak termasuk data yang dipakai dalam fase pembelajaran metode itu. *SVM* tidak dipengaruhi oleh dimensi dari input vector, sehingga SVM merupakan salah satu metode yang tepat dipakai untuk memecahkan masalah berdimensi tinggi, dalam keterbatasan sampel data yang ada. (Nugroho, Witarto, & Handoko, 2003)

Berdasarkan uraian diatas maka penulis tertarik untuk melakukan penelitian mengenai klasifikasi kategori berita menggunakan metode *SVM* dikombinasikan dengan reduksi fitur menggunakan *SVD*. *SVD* digunakan untuk mengurangi fitur yang sangat besar hasil dari proses *TF-IDF* serta menjaga kinerja dari sistem.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, maka rumusan masalah yang akan dibahas adalah sebagai berikut:

a. Bagaimana penerapan seleksi fitur untuk klasifikasi kategori berita?
b. Bagaimana penerapan metode *SVM* untuk klasifikasi kategori berita?
c. Bagaimana performa metode *SVM* dalam melakukan klasifikasi kategori berita?

1.3 Batasan Masalah

Adapun batasan masalah yang digunakan peneliti agar pembahasan dalam penelitian ini tidak menyimpang dari pembahasan adalah sebagai berikut:

a. Data yang digunakan adalah berita berbahasa Indonesia dan berasal dari situs kompas.com.
b. Data yang digunakan telah didefinisikan oleh penulis dalam *database.*
c. Metode klasifikasi yang digunakan adalah *SVM* dengan *kernel Polynomial.*
d. Bahasa pemrograman yang dipakai untuk klasifikasi adalah JAVA.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang dirumuskan sebelumnya, maka tujuan dari penelitian ini dapat dituliskan antara lain:

a. Mengimplementasikan metode *SVM* untuk mengkategorikan berita.
b. Mengetahui performa dari metode *SVM* untuk mengkategorikan berita.
c. Menginplementasikan *SVD* sebagai reduksi fitur dalam proses klasifikasi kategori berita.

1.5 Manfaat Penelitian

Penelitian ini dilakukan dengan harapan dapat memberi sebuah manfaat yaitu membantu editor berita dalam melakukan pengkategorian terhadap berita yang akan di terbitkan menjadi lebih efesien karena kategori berita akan dimasukkan secara otomatis dengan komputer menggunakan metode tertentu.