BAB I

Pendahuluan

1.1 Latar Belakang

Pesatnya perkembangan internet di dunia menjadikan penjualan barang *secara online* semakin banyak diminati. Banyak situs *e-commerce* dan *market place* saat ini menyediakan penjualannya di internet, salah satunya adalah *amazon.com, Inc. Amazon.com, Inc*. merupakan perusahaan yang bergerak dalam bidang perdagangan berbasis daring yang berpusat di *Seattle*, Amerika Serikat. Produk elektronik merupakan salah satu produk yang memiliki banyak *review*. *Review* ini menimbulkan munculnya berbagai sentimen tentang kepuasan konsumen terhadap barang yang dibeli. Sentimen konsumen ini sangat penting dalam suatu pengambilan keputusan bagi konsumen lain untuk mengetahui kualitas suatu produk yang akan dibeli. Di sinilah peran analisis sentimen diterapkan untuk menentukan ulasan yang bernilai postif atau negatif.

Banyak algoritma klasifikasi dengan pendekatan *supervised machine learning* diterapkan dalam klasifikasi teks salah satunya adalah *Naïve Bayes Classifier* (NBC). *Naïve Bayes Classifier* (NBC) adalah teknik pembelajaran untuk *text classification* dan memiliki performa baik di berbagai domain (Qiang Ye, Ziqiong Zhang, Rob Law, 2009). Tapi, terlepas dari performanya yang baik *Naïve Bayes Classifier* sangat sensitif terhadap pemilihan fitur. (J. Chen, et al, 2009).

Selain itu semakin meningkatnya volume teks juga akan menimbulkan banyaknya atribut yang dipakai dalam klasifikasi dan akan mengurangi kinerja algoritma klasifikasi (J. Ipmawati, Kusrini, Emha T. L., 2017). Maka dari itu teknik *preprocessing* data penting dalam klasifikasi sentimen (I. Hemalatha, G. P Saradhi, A. Gofardhan, 2012).

Feature selection merupakan teknik pemilihan fitur, menurut (Vinita Chandani, Romi Satria W, Purwanto, 2015) *feature seletion* dapat digunakan untuk menyeleksi banyaknya fitur yang tidak relevan pada dataset. Beberapa algoritma *feature selection* telah diterapkan seperti *Mutual Information* (IG), *Term Frequency* (TF), *Information Gain* (IG), *Chi-Square*, *Expected Cross Entropy*, *Odds Ratio*, the *Weight of Evidence*, dan *Gini Index* banyak digunakan dan memiliki performa yang baik (Vandana Korde, C. Namrata M, 2012). Berdasarkan penelitian (J. Ling, N. E. I.Kencana, T. B. Oka, 2013) penggunaan *Chi-Square* disarankan dengan menggabungkan algoritma klasifikasi *Naïve Bayes Classifier* dan berdasarkan (Novan D. P., Yuita A. S., Putra P. A., 2018) penggunaan model *N-Gram* disarankan untuk membantu dalam penyeleksian fitur dan juga untuk memeriksa keakuratan hasil klasifikasi yang diperoleh. Untuk itu penelitian ini mengajukan pendekatan dengan menggunakan metode *Naïve Bayes Classifier* serta penggunaan *Chi-Square* sebagai *feature selection* dengan mengkomparasikan *N-Gram* (*uni-gram*, *bi-gram*, dan gabungan *uni dan bi-gram*).

1.2 Rumusan Masalah

Berdasarkan permasalahan yang ada dapat dirumuskan, bagaimana pengaruh *feature selection* menggunakan *Chi-Square* dan *N-Gram* terhadap tingkat akurasi, presisi, dan *recall* pada sentimen *review* menggunakan metode *Naïve Bayes Classifier*?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengetahui pengaruh *feature selection* menggunakan *Chi-Square* dan *N-Gram* dalam peningkatan akurasi, presisi, dan *recall* terhadap metode *Naïve Bayes Classifier* pada sentimen *review* produk elektronik.

1.4 Manfaat Penelitian

Penerapan algoritma klasifikasi *Naïve Bayes Classifier* dengan pemilihan fitur *chi-square* dan *N-Gram* dapat memberikan informasi tentang keakuratan hasil dalam menentukan ulasan yang bernilai positif dan negatif tentang ulasan produk secara *real time* sehingga dapat dipergunakan sebagai pengambilan keputusan bagi konsumen lain untuk mengetahui kualitas suatu produk yang akan dibeli.

1.5 Batasan Masalah

Adapun batasan-batasan dalam penelitian ini.

1. Dataset merupakan *review* produk elektronik situs *www.amazon.com* yang di unduh dari www.*kaggle.com* dengan jumlah dataset yang digunakan 600 *review* teks Bahasa Inggris.

2. Dataset yang di olah meliputi data *review* produk elektronik yang memiliki *rating* 1 (satu), 2 (dua), 4 (empat), dan 5 (bintang).
3. *Prepocessing* hanya diterapkan pada teks Bahasa Inggris.
4. *Feature selection* menggunakan *Chi-Square* dan *tokenizing* menggunakan metode *N-Gram* dengan teknik *uni-gram*, *bi-gram*, dan *uni+bi-gram*.
5. Klasifikasi menggunakan metode *Naïve Bayes Classifier*.
6. *Package R* program yang digunakan adalah *tm 0.7-1, RWeka 0.4-39, e1071 1.6-8, FSelector 0.31, caret 6.0-81, SnowballC 0.6.0, ggplot2 3.1*.

UNIVERSITAS MUHAMMADIYAH JEMBER