ABSTRACT

Dissemination of negative content is a problem that comes with the rapid development of the internet. One of the efforts that have been made to overcome this problem is the use of a negative content filtering system based on classification of text and image contents. Text classification plays an important role in recognizing negative content on the internet. In the text classification process, high dimensionality vector and loss of semantic relationships between features are issues which become concern in several studies.

In this work, selective extraction of n-gram features and class information-based feature weighting methods are proposed to be alternative methods to deal with both issues above. Selective extraction of n-gram features method attempts to extract n-gram features selectively to avoid the first issue. Class information-based feature weighting method uses available information in the classification process to resolve the second problem.

Testing of both methods carried out using CFI data, 10-fold cross validation, and statistical significance test. The test results revealed that selective extraction of n-gram features method significantly affects the classification results. Meanwhile, the use of class information-based feature weighting has no significant effect on the classification results.

Keywords: classification, feature extraction, n-gram, features, feature weighting.

INTISARI

Penyebarluasan konten negatif merupakan masalah yang muncul seiring berkembang pesatnya internet. Salah satu upaya yang telah dilakukan untuk mengatasi masalah ini adalah penggunaan sistem penapis konten negatif berbasis klasifikasi konten teks dan gambar. Klasifikasi teks berperan penting dalam pengenalan konten negatif di internet. Pada proses klasifikasi teks, masalah *high dimensionality vector* dan *loss of semantic relationship* antar fitur menjadi perhatian pada beberapa penelitian.

Pada karya ini, metode ekstraksi selektif fitur n-gram dan pembobot fitur berbasis informasi kelas diajukan untuk menjadi metode alternatif dalam menghadapi kedua masalah di atas. Metode ekstraksi selektif fitur n-gram berupaya mengekstrak fitur n-gram secara selektif untuk menghindari masalah pertama. Metode pembobot fitur berbasis informasi kelas memanfaatkan informasi yang tersedia dalam proses klasifikasi untuk mengatasi masalah kedua.

Pengujian kedua metode dilakukan menggunakan data CFI, validasi 10-fold, dan uji signifikansi statistik. Hasil pengujian menunjukkan metode ekstraksi selektif fitur n-gram secara signifikan berpengaruh pada hasil klasifikasi. Sementara itu, penggunaan metode pembobot fitur berbasis informasi kelas tidak berpengaruh signifikan pada hasil klasifikasi.

Kata kunci: klasifikasi, ekstraksi fitur, n-gram, fitur, pembobot fitur.