ABSTRAK

Web *harvesting* dari sebuah mesin web *crawler* untuk artikel kesehatan berbahasa Indonesia dapat menjadi sumber informasi kesehatan yang *cost-effective* di Indonesia. Agar dapat dimanfaatkan secara optimal seperti untuk *machine learning*, maka beberapa tahapan harus dilakukan seperti *text pre-processing* dan *clustering* berdasarkan topiknya. Pada penelitian ini metode yang digunakan terbagi menjadi empat tahapan. Tahapan pertama yaitu *text pre-processing* yang terdiri dari *tokenizing, case-folding, filtering, phrase detection,* dan *stemming*. Tahapan kedua yaitu pembobotan kata dari setiap artikel yang ada dengan menggunakan metode TFIDF. Tahapan ketiga yaitu ekstraksi kata kunci dari setiap artikel. Tahapan terakhir yaitu proses *clustering* dengan algoritma *Self Organizing Maps* yang dibagi menjadi dua proses. Proses pertama dari *clustering* adalah memisahkan artikel kesehatan ke dalam dua korpus yaitu artikel kesehatan yang berhubungan dengan anak dan artikel kesehatan umum. Proses kedua *clustering* yaitu mengelompokkan artikel kesehatan pada setiap korpus tersebut berdasarkan topiknya. Pada penelitian ini jumlah artikel yang di-*cluster* adalah sebanyak 533 artikel. Hasil dari penelitian ini adalah dua korpus baru yang berisi *cluster-cluster* sesuai dengan topik pada setiap korpus. Kemudian artikel-artikel yang memiliki kesamaan *term* akan berada pada *cluster* yang sama.

Kata kunci: *Text mining, Clustering,* TFIDF, *Self Organizing Maps, Multiword Expression*

HEALTH WEB ARTICLE CLUSTERING WITH SELF ORGANIZING MAPS ALGORITHM

ABSTRACT

Web harvesting from a web crawler machine for Indonesian health article can be a cost-effective health information source in Indonesia. In order to be optimally used for such machine learning, then some processes like text pre-processing and clustering have to be done. The method used in this research divided into four steps. First step is text pre-processing that consist of tokenizing, case folding, filtering, phrase detection, and stemming. Second step is term weighting for all terms in all articles in corpus using TFIDF method. Third step is keyphrase extraction from each article. The last step is clustering with self organizing map algorithm, this step divided by two processes. The first process of this clustering was to seperate the articles into health article related to children corpus and general health article corpus. The second process of this clustering was to categorize article each of corpuses based on its topic. In this reseach, the number of articles are clustered is 533 articles. The results of this study are two new corpuses containing clusters according to the topics in each corpus. Then the articles that have similar terms will be on the same cluster

Keywords: *Text mining, Clustering,* TFIDF, *Self Organizing Maps, Multiword Expression*