ANALISIS SENTIMEN DATA *TWITTER* MENGGUNAKAN *K-MEANS CLUSTERING*

Gregorius Agung Purwanto Nugroho

ABSTRAK

Penelitian ini bertujuan untuk menciptakan sistem untuk mengenali emosi yang terkandung dalam kalimat *tweet*. Latar belakang penelitian ini yaitu maraknya penggunaan media sosial atau *microblogging* untuk mengutarakan opini tentang topik tertentu. Penelitian berkaitan opini publik dapat dijadikan sebagai dasar manajemen merk, *corporate reputation*, marketing, sistem rekomendasi, dan intelijen.

Penelitian ini menggunakan metode *K-Means Clustering* dengan masukan berupa teks. Penelitian mencakup tahap *preprocessing*, pembobotan, normalisasi, *clustering*, dan uji akurasi. *Peprocessing* meliputi *tokenizing*, *remove stopword*, dan *stemming*. Pembobotan menggunakan metode *term frequency-inverse document frequency (tf-idf)*. Normalisasi menggunakan *z-score* dan *min-max*. Clustering menggunakan *K-Means* dengan penentuan *centroid* awal memakai *Variance Initialization* dan hitung kemiripan dengan *Cosine Similarity*. Pengujian akurasi memakai metode *Confusion Matrix*.

Percobaan dilakukan pada 1000 data yang dikelompokkan menjadi lima *cluster* yaitu cinta, marah, sedih, senang, dan takut. Akurasi tertinggi sebesar 76,3%. Hasil akurasi tertinggi didapat dengan metode normalisasi *min-max*, batas nilai yang dinormalisasi 5, dan minimal kemunculan kata 3.

Kata Kunci: *Tweet, K-Means Clustering, Cluster, Centroid, Variance Initialization, Cosine Similarity, Confusion Matrix*

SENTIMENT ANALYSIS OF TWITTER DATA USING K-MEANS CLUSTERING

Gregorius Agung Purwanto Nugroho

ABSTRACT

The objective of this research is to create system to recognize emotion of a tweet. This research is created to learn public opinion about a certain topic. The study about public opinion can be used as the key factor to determine brand management, corporate reputation, marketing, recommendation system, and intelligent.

The research uses the K-Means Clustering as the main algorithm and textual data as the input. The research includes the preprocessing, the weighting, the normalization, the clustering, and the accuration testing. The preprocessing includes the tokenizing, the stopword removal, and the stemming. The weighting uses the term frequency - inverse document frequency (tf-idf) method. The normalization uses the z-score and the min-max method. Clustering uses the K-Means Clustering with the Variance Initialization method to determine the initial centroids and the Cosine Similarity method to measure the similarities. The testing uses the Confusion Matrix.

The experiment has been applied to a data sets of 1000 tweets that divided into five clusters: *cinta* (love), *marah* (anger), *sedih* (sadness), *senang* (happiness), and *takut* (fear). The experiment obtained the highest accuration of 76.3% using the min-max normalization, the min-max threshold was 5, and the minimum word frequency was 3.

Keywords: Tweets, K-Means Clustering, Clusters, Centroids, Variance Initialization, Cosine Similarity, Confusion Matrix