Research Paper 2: Opinion Mining on E-Commerce Data Using Sentiment Analysis And K-Medoid Clustering

Abstract:

The method used was through analyzing text reviews obtained from customers on an e-commerce website. The algorithm used was k-medoid clustering.

Introduction:

- Several previous studies have explained, that in conducting product evaluation methods are used based on star rating, where customers must provide ratings from one to five according to the quality of the product.
- However, sometimes the rating is not in accordance with the product reviews given in textual form. Many occur on several e-commerce websites which show that a product has a 4- or 5-star product rating, but in the review column there are many negative opinions, which make information on the quality of the product biased. So, the use of text mining concepts is important to implement.
- The purpose of this study is to sort out unstructured words into information so that they can distinguish between positive and negative opinions.

Related Work:

- The purpose of classifying sentiments or opinions is to identify and do the polarity of various kinds of opinions.
- There are two models of analytical sentiment approaches, namely lexicon-based and machine learning.
- In the lexicon-based analysis sentiment the method of calculating on documents uses the semantics of words or phrases.
- But in machine learning data that is trained is used to group opinions or sentiments.
- Some methods that can be used in machine learning are naive bayes, k-means, k-medoid and support vector machine (SVM)
- K-medoid is a clustering algorithm similar to K-Means. The difference between the two algorithms is that the k medoid algorithm uses objects as representatives (medoid) as the center of the cluster for each cluster, while K-Means uses the mean as a center
- The k-medoid algorithm has the advantage of overcoming the weaknesses of the K-Means algorithm which is sensitive to noise and outliers
- Another advantage is that the results of the clustering process do not depend on the order in which the dataset is entered

Proposed Approach:

- The main architecture of the approach, the process consists of two phases. The first pre-processing phase consists of tokenize, stop words filtering, transform cases, stemming, and TF-IDF. Second phase is clustering with k-medoid.

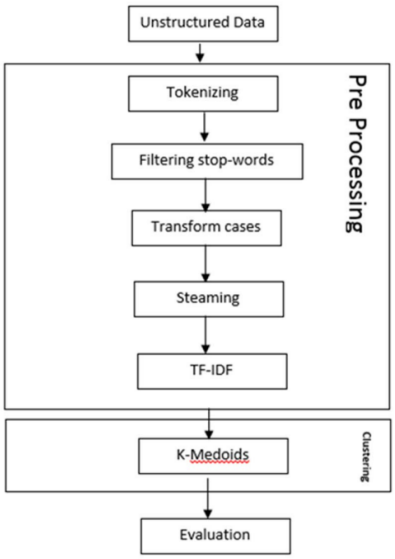

Figure 1. Architecture of the proposed scheme

A. Data Collection

The data used in this study amounted to 88 data. The comment data is obtained from the site tokopedia.com and shopie.com, then the data is processed using the Python programming language.

B. Preprocessing Data:

that has been collected is then processed in preprocessing, before going to the data clustering stage. In this study there are several preprocessing stages used, namely: Tokenization, Filtering stop words, transform cases (The words that have been obtained are then changed to all lowercase letters.), Steaming, Weighting words (TF-IDF).

The formulas for these statistics are provided below.

$$TFIDF_{ij} = TFij \times idf_i$$

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

$$IDF_i = \log \frac{|D|}{\left|\{j : t_i \in d_j\}\right|}$$

Here n_{ij} represents the appearance of the word i in the document d_j. $\sum_k n_{kj}$ is the number of total words which appear in the document. $|D|$ is the total number of words in the corpus, and $\left|\{j : t_i \in d_j\}\right|$ is the number of corpora which includes the word t_i.

C. Clustering:

- the k-medoid algorithm uses actual data points to represent clusters. The algorithm is less sensitive to outliers than k-means.
- Step 1.
  Initialization: randomly select k from n data objects as medoid.
  Step 2.
  Link each data object to the closest medoid m, which is defined using the formula distance between the Sentiment Sequence above.
  Step 3.
  Look for a new medoid from each cluster, which is an object that minimizes the cost function, which is the total distance to other objects in the cluster. Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is_
- Cost Function: $$J = \sum_{n=1}^{N} \sum_{k=1}^{K} D(X_n, \mu_k)$$

  Step 4.
  Repeat step 2 to 4 until there are no changes to the medoid. The k-medoid method can divide the documents specified into more than three categories, it is done well, efficiently, and non-human participation. In this study the cluster was divided into 2 parts, to divide the comment data into positive clusters and negative clusters.

Results:

Before analyzing the performance of the k medoid method in grouping comments into 2 groups, Figure 2 shows the percentage and number of each cluster, positive and negative.

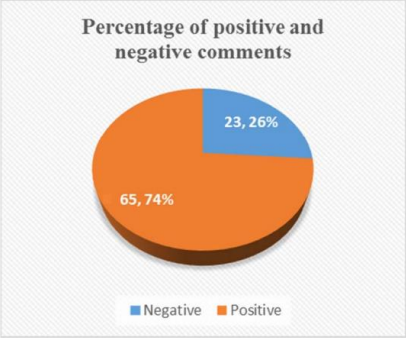

Figure 2. Percentage of positive and negative comments

To find out whether the cluster results obtained were good or not, we measured the performance of k-medoid with a mean in the cluster distance (the average distance between the cluster centroid and all elements in a cluster) and the Davies-Bouldin Index (DB) (which is given as a cluster, this metric measures the average similarity between each cluster and the most similar.)

TABLE 1. K-MEDOID ALGORITHM PERFORMANCE

Avg. within centroid distance	Davies Bouldin
1.778	1.867

CONCUSION:

- In our approach, sentiment analysis is applied to the phrase level rather than the document level to calculate the polarity of sentiments for each term.
- Then key graph keyword extraction techniques are used to extract keywords from each document with the term high frequency.
- After founding a list of the most used words through TF-IDF. Then we calculated the intensity of the sentiment polarity by measuring its strength.
- To summarize and minimize the purpose of the data analyzed, we use k-medoid algorithm to group and group data based on sentiment strength values.
- We also compare the results of our techniques with the same star rating data and find a better sentiment towards the product.
- As a star rating gives a biased result; by analyzing the review text through this approach the sentiments of each term can be obtained.