Abstract

Product reviews contain rich user opinions on quality products. They are used by potential customers to find opinions of existing users before deciding to purchase a product. The product reviews also used by product manufacturers to identify product problems and to find marketing intelligence information about their competitors.

However, because there is no quality control on writing reviews that can be trusted, then everyone is free to give his opinion that it is not uncommon opportunity exploited by spammers. There are three types of spam on product reviews are untruthfull opinions, reviews on the brand only, and non-review.

The method chosen to solve this case is the Logistic Regression because it can produce estimates of the probability of a review is spam or not as required. In this case, the whole system is divided into two system, first for classification to spam 1 and the second is classification to spam 2 and 3. For classification spam 2 and 3 will be done by manually labeling first, then do the preprocessing to find the desired predictive variables. Due to the spam 1 classification, manual labeling is not done, so it will be separated of review duplicate and non-duplicate by the shingling method first. In which later the duplicate review will be considered as spam and the rest as non-spam. And then, it would be analyzed the characteristics of spam type 1. From the analysis, we obtained that characteristics of spam 1 became a target of spammers are from review which have a good or average quality product (rating 3-5), an then they will give a negative review spam to drop the product.

Key words : Logistic regression, classification, duplicate, spam