**3.3.1 Detection of Duplicate Reviews**

* This system uses **duplicate reviews** as one of the marks to score the credibility of the product.
* One of the main challenges we are solve is differentiating spam vs. non-spam reviews in the Amazon dataset.
* We will define four types of duplicate reviews:
  + duplicates from different reviewer IDs on the same product,
  + duplicates from different reviewer IDs on different products,
  + duplicates from the same reviewer ID on the same product,
  + duplicates from the same reviewer ID on different products.

Figure 4 shows the steps used in detecting the duplicate reviews.

Diagram

Description automatically generated

To have computational efficiency for a very large dataset, the following optimizations were done in each step.

1. Each review is converted into a set of bigram shingles, which are formed by combining two consecutive words together.
2. The shingles are mapped to shingle IDs using the CRC32 hash. These ‘min-Hash’ signatures were calculated for each review using a random hash function which prevents from having to explicitly compute random permutations of all the shingle IDs.
3. Then, the Jaccard similarities are calculated using the min-Hash signature for each review. They are compared by counting the number of equal signatures divided by the number of matching components by the signature length to get a similarity value. The inverted index was used to compute similarity faster. The output displays pairs of reviews with similarity greater than a set threshold. In this case, review pairs with a similarity score of at least 70% were chosen as duplicates [22]. The following sections from 3.3.1.1 to 3.3.1.7 explain each of these steps in detail.

Figure 4: Overall steps in Detection of Duplicate Reviews.

29

3.3.1.1 Jaccard Similarity

In general, the similarity is the degree of alikeness between two data objects. In data mining, it is the distance between the object’s features. The similarity is measured in the range of 0 to 1. If the similarity is equal or closer to 1, the distance is small and signifies a high degree of similarity between the two objects. If the similarity is equal or closer to 0, the distance is large and signifies a low degree of similarity between the two objects [6]. There are vari- ous similarity measures like Euclidean distance, Cosine similarity, Minkowski distance, and Jaccard similarity. For this research, Jaccard similarity was used to calculate the similarity between two texts of the reviews. Other similarity measures are used if the data objects are represented as vectors or points. Jaccard similarity is used for data objects represented as sets (an unordered collection of objects).

Jaccard similarity is defined as the ratio of the cardinality of the intersection of sets over the cardinality of the union of the sets [6]. The resulting score indicates the amount of sim- ilarity between the two sets. For any two sets A and B, Jaccard similarity is represented by the formula: