3 Proposed Duplicate Removal Technique

A simple and new approach has been used to design the data cleaning system, especially for duplicate detection and removal which is shown in Fig. 1.

3.1 Duplicate Removal Algorithm

The data from the text file is read word by word and stored in the array list. The length of the array is calculated. In the process, each word/character in the array is compared with all other words/characters to detect the duplicates. If the duplicate is found, it is deleted and the count is maintained for the repeated number of words/characters. The modified array along with the duplicate information is saved.

- Read the data from the file and store it into the ArrayList.
- Calculate the length of ArrayList.
- Initialize variable i, j, and c with 0.
- Check the length of ArrayList
  - If length > i, go to step *v,* else
  - store the original ArrayList in file and go to step xi.
- $y = i + 1$
- While (j < length-1), go to step vii, else go to step x.
- Compare value of $a[i]$ with $a[j]$, if yes go to step viii, else increment j by 1 and go to step vi.
- Add duplicate element in new ArrayList and Delete duplicate element from the original ArrayList. Count the duplicate element by incrementing the value of c by 1.
- Increment value of j by 1.
- Calculate the new length of updated ArrayList and increment the value of i by 1 and goto step iv.
- End.

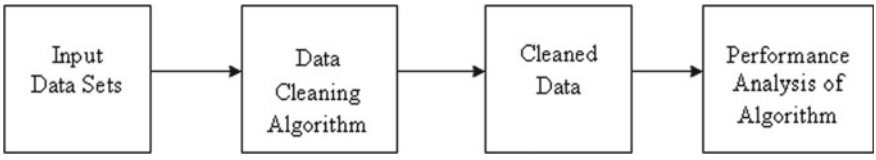

Fig. 1 Block diagram of the system