Table 1 Performance analysis on varying numbers of words

File size	Number of duplicate words	Number of words after removing duplicates	Duplicate detection rate (%)
10	03	07	100
25	07	18	100
65	13	52	100
95	27	68	100
133	35	98	100
356	148	208	100
578	189	389	100
753	256	497	100
1243	365	878	100
1986	487	1499	100
2576	583	1993	100
3563	654	2909	100
4587	666	3921	100
5643	743	4900	100
6487	798	5689	100
7489	1008	6481	100
8734	1154	7580	100
9679	1678	8001	100
18,765	4563	14,202	100
23,598	5003	18,595	100

5.5 Performance Analysis on Varying Numbers of Total Words

To test the algorithm on a file that contained huge number of data, different size files are tested. The results are excellent as expected and shown in Table 1.

6 Conclusion

A data cleaning algorithm is presented in order to detect and remove duplicate data from text files having different data types in order to improve the quality of data. From the results, it could be found that the proposed algorithm could provide 100% correct duplicate detection rate in all the cases. The proposed concept can be extended for removal of other kinds of data impurities. This work presents testing with text files only, and hence, the algorithm can be modified for duplicate removal from other file formats such as xls, xlsx, doc, docx, and pdf.