Ujjwala-Textual and Demographic De-duplication: Facilitating the PMUY

The project to de-duplicate the textual and demographic data for the Pradhan Mantri Ujjwala Yojana (PMUY) has enabled elimination of many fraudulent beneficiaries and holds immense potential for application in a variety of other schemes initiated by the Government.

G. MAYIL MUTHU KUMARAN
Technical Director
muthu@nic.in

DEEPJOT KAUR
Scientist- B
deepjot.kaur@nic.in

Edited by
MOHAN DAS VISWAM

India is widely touted as an emerging economic power, but nearly half of the country still cooks with firewood. According to 2011 census, only 28.5% of the 1.3 billion population across India uses LPG/ PNG for cooking. Pradhan Mantri Ujjwala Yojana (PMUY) aims at providing 5 crore free LPG connections in 3 financial years (2016-19) to the women belonging to below poverty line (BPL) [as per the Socio Economic Caste Census data] across the country.

The Government extends support to the population in the form of various schemes, subsidized policies and grants. However, in governance, enormous data of individuals gets collected in digitized databases. At times certain duplicates infiltrate the digitized databases which pave the way for misuse of the government subsidies by unintended and fraudulent beneficiaries.

Identifying the duplicates from a large chunk of applications become a tedious task. Developing a de-duplication algorithm requires grit and thorough assessment of raw data. Textual and Demographic De-duplication (TDD) algorithm aims to identify transformed or direct forms of same person's identity exhibited as multiple persons in the database. The results are thus called as SUSPECTS. The SUSPECTS found are sent for field verification to the department. If the SUSPECT turns out to be a real deceitful case, then it is blocked from the benefited list and is termed as a DUPLICATE, else it is rejected as a FALSE SUSPECT. In essence, TDD is developed with an intention to use technology in the form of algorithm without the use of any biometrics, expensive equipment or beneficiary's presence to excerpt the fraudulent suspects. Till date, TDD has successfully de-duplicated

> “I would like to convey my appreciation for the work done by the team working on de-duplication of LPG beneficiaries for the Oil Marketing Companies on behalf of Ministry of Petroleum and Natural Gas. The team has contributed immensely towards the implementation of Pradhan Mantri Ujjwala Yojana (PMUY) scheme by extracting the 10.15 Crore beneficiaries' dataset from the SECC (Socio Economic Caste Census) data and de-duplicating the applicants in a time bound manner. Their sincere efforts in de-duplicating Kerosene Oil user's data against the LPG beneficiary database also deserves to be commended.”
>
> **ASHUTOSH JINDAL, IAS**
> Joint Secretary (Marketing)
> Ministry of Petroleum & Natural Gas

15.5 crore LPG beneficiaries curtailing around 1,56,33,099 multiple gas connections and is continued henceforth under PAHAL scheme to aid in saving government funds and mitigating frauds. The same algorithm with revised logic is now used for de-duplicating PMUY data.

PMUY DE-DUPLICATION

There is no straight forward approach to deal with a large chunk of miscellaneous data. A subtle understanding of the scenario is required to formulate the process according to the requisite. The working scenario of PMUY is as follows:

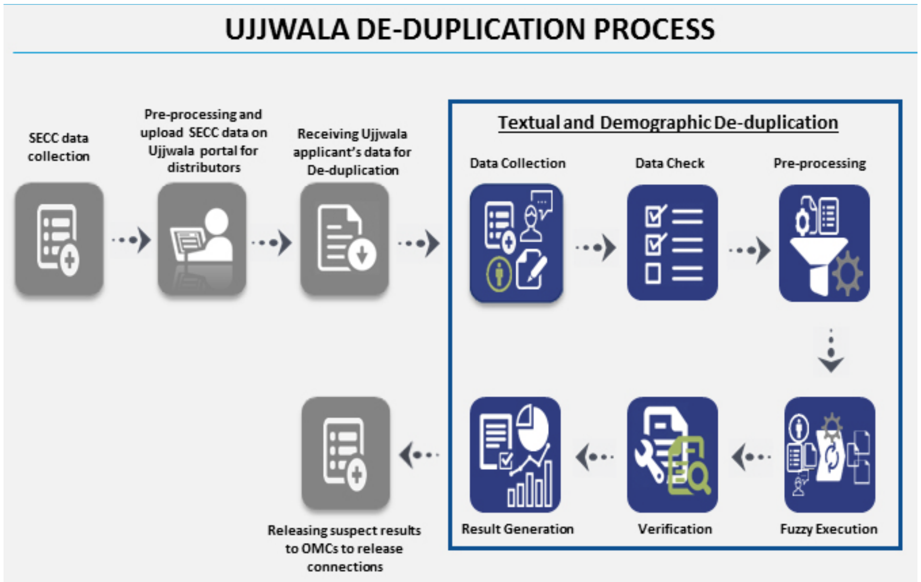
1. Receiving Socio Economic Caste Census (SECC) data from TDP (Telematics Development Promotion) division.

2. Preparing the data to be in a unified standard format and manifesting the data to OMCs (Oil Marketing Companies) through which OMCs check the eligibility criteria of the applicant, applying for a new gas connection under PMUY scheme.

3. Receiving the PMUY applicant data from OMCs as Waitlist data and de-duplicating the data on the basis of three textual demographic parameters – Name, Address and KYC (Know Your Customer) number; which is AHL-TIN (Abridged Household List - Transaction Identification Number) which is a 29 digit unique number. The first 27 digits of the AHLTIN are same for each family and the last 2 digits differ for each member within the family. De-duplication is done for both family and members of the family so as to ensure there are no duplicates within the family.

4. Forwarding the clear list (Beneficiaries who are not suspects and qualified to receive the connection under PMUY) to OMCs through Web Services for releasing the connection to the beneficiaries.

5. The data of the beneficiaries who got the new connection through PMUY will be sent again to NIC as delta data for demographic de-duplication.


UJJWALA DE-DUPLICATION PROCESS
SECC data collection
Pre-processing and upload SECC data on Ujjwala portal for distributors
Receiving Ujjwala applicant's data for De-duplication
Textual and Demographic De-duplication
Data Collection
Data Check
Pre-processing
Fuzzy Execution
Verification
Result Generation
Releasing suspect results to OMCs to release connections

Ujjwala De-duplication Architecture

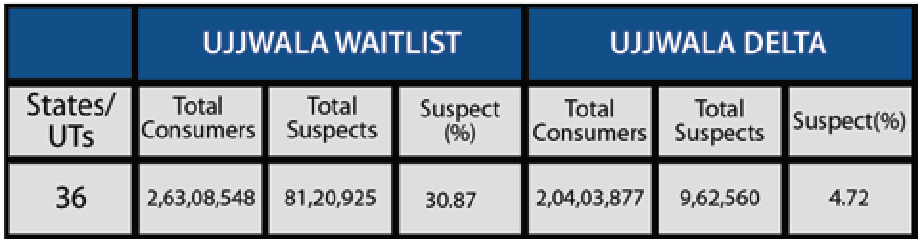
ROLE OF DE-DUPLICATION IN PRADHAN MANTRI UJJWALA YOJANA (PMUY)

NIC's role in curtailing the duplication in PMUY is immense and requires collection and analysis of large set of data constructively. Be it collecting SECC data from TDP, unifying the data and making it available to OMCs for validating the applicants' eligibility or be it receiving PMUY Waitlist/ Delta data and de-duplicating, NIC provides its methodologies in executing the whole process in a well-defined manner.

De-duplication for PMUY waitlist data kick-started in April 2016 with all 36 States/ UTs going live. The data of the beneficiaries who have received a new connection through PMUY are sent to NIC as delta data for demographic de-duplication to refurbish the existing beneficiary data. De-duplication for PMUY Delta data commenced from January 2017. Real time de-duplication for Delta data is being performed for all the States/ UTs on daily basis.

WAY FORWARD

Central and State Governments offer numerous welfare schemes for the people, but ensuring their accessibility to the intended beneficiaries is still a far cry. In this context, curtailing the duplicates without the use of sophisticated biometrics infrastructure is one of the major issues faced. One area where Textual and Demographic De-duplication can be of value is in ensuring fair usage of government funds in projects like Public distribution system, agricultural subsidies, petroleum subsidies, educational grants, development authority subsidies and other such sectors. Another area is preventing the misuse of Government identification cards where one individual possess multiple/ duplicate identification cards. To name a few passports, government health cards, election IDs, driving licenses, insurance policies and other similar domains. All these are sectors where textual and demographic de-duplication has immense application.



	UJJWALA WAITLIST			UJJWALA DELTA		
States/ UTs	Total Consumers	Total Suspects	Suspect (%)	Total Consumers	Total Suspects	Suspect(%)
36	2,63,08,548	81,20,925	30.87	2,04,03,877	9,62,560	4.72

For further information, please contact:

G. MAYIL MUTHU KUMARAN
Technical Director
NIC HQ, A- Block, CGO Complex
Lodhi Road, NEW DELHI- 110 003
Phone: 011-24305748
Email: muthu@nic.in