

READ THIS FIRST

What this file is. A sample project proposal (synopsis) for IGNOU MCSP-079 Project Work, M.Sc. (Data Science and Analytics), written by ignouproject.com as study reference. It shows how the five proposal elements named in the MSCDSA Programme Guide can be developed on one realistic topic.

What it is not. It is not a synopsis to submit. Do not copy it, change the name and send it. IGNOU expects the learner's own original work; copied work can be treated as unfair means.

What the official documents say (read 06 October 2026). The MSCDSA Programme Guide (SOCIS, January 2026, version 1) asks the learner to begin with a proposal that outlines the problem statement, objectives, methodology, expected outcomes and deliverables, followed by a literature review, data collection and preprocessing, machine learning or statistical analysis, model evaluation and interpretation. It says the proposal is submitted during the third semester through the Regional Centre or Learner Support Centre, and that detailed project guidelines will be made available on the University website. As of 06 October 2026 no separate MCSP-079 guideline or approval proforma had been published, so this sample sets no page count, font rule or proforma. Follow the MCSP-079 guideline and proforma as soon as IGNOU releases them.

How to use the brackets. Everything in [square brackets] is a fact only you can supply: your dataset, its source and permission, your organisation, guide, dates and numbers. No data, result or statistic in this file is real, and none should be invented.

Need your own draft checked against the guide before you submit it? ignouproject.com reviews and corrects learner-written synopses; you remain the author and IGNOU's evaluator decides approval.

ignouproject.com

PROJECT PROPOSAL (SYNOPSIS)

MCSP-079: Project Work

Detecting Near-Duplicate Public Complaints Using MinHash Locality-Sensitive Hashing and Sentence Embeddings: A Comparative Study on [Name of Complaint Dataset]

Submitted by: [Learner's full name]

Enrolment No.: []

Programme: Master of Science (Data Science and Analytics) (MSCDSA)

Regional Centre: [] Learner Support Centre: []

Proposed Guide: [Name, designation, organisation]

Organisation where the work will be done (if any): []

School of Computer and Information Sciences, IGNOU

ignouproject.com

1. Introduction and Problem Statement

Public grievance systems, utility helpdesks and customer-care portals receive large volumes of short free-text complaints. Many of them describe the same underlying issue in different words: the same broken streetlight reported by several residents, or the same billing error raised again after a reply. When near-duplicates are not grouped, the same issue is counted several times, staff effort is repeated and the true volume of distinct problems is hard to see.

Problem statement. The study will examine how accurately and how efficiently near-duplicate complaints in [name of complaint dataset, with its source] can be identified by (a) a lexical baseline using TF-IDF vectors and cosine similarity, (b) MinHash signatures with locality-sensitive hashing (LSH) on word shingles, and (c) sentence embeddings from a pre-trained transformer model, and what each approach costs in computation as the number of complaints grows.

The topic draws directly on the MSCDSA syllabus: similarity measures, Jaccard similarity, shingles, min-hashing and LSH (MCS-068, Block 2, Unit 5), performance evaluation measures (MCS-068, Unit 12), transformers and information retrieval (MCS-077), Spark for large-scale processing (MCS-071) and data security (MCS-078).

ignouproject.com

2. Literature Review and Gap

Broder (1997) introduced the idea of measuring the resemblance of documents through shingling and min-wise permutations, which underlies MinHash. Indyk and Motwani (1998) formalised locality-sensitive hashing for approximate nearest-neighbour search, the basis for comparing items without checking every pair. Leskovec, Rajaraman and Ullman (2020) present shingling, MinHash and LSH together as a practical pipeline for finding similar items at scale. On the semantic side, Devlin et al. (2019) introduced BERT, and Reimers and Gurevych (2019) adapted it into Sentence-BERT so that sentence embeddings can be compared efficiently with cosine similarity.

[Add two to four recent studies you have actually read on duplicate detection in complaint, ticket or question-answer data, with full citations, and state in one or two sentences what each found and on what data.]

Gap to be addressed. [State, from the studies you read, what is not yet known for your data. A typical gap: lexical hashing methods and embedding methods are usually evaluated on long documents or English question pairs; evidence is limited for short, informal, possibly code-mixed complaint text of the kind found in [dataset].]

ignouproject.com

3. Objectives

The objectives are worded neutrally and do not assume which method will perform better.

1. To prepare a cleaned, de-identified corpus of complaints from [dataset] and a hand-labelled sample of complaint pairs marked as duplicate or not duplicate.
2. To implement three near-duplicate detection approaches: a TF-IDF cosine baseline, MinHash with LSH on word shingles, and sentence-embedding similarity.
3. To compare the three approaches on precision, recall and F1-score for the duplicate class, and on running time and memory as corpus size increases.
4. To examine how the similarity threshold and, for LSH, the number of bands and rows affect the balance between missed duplicates and false matches.
5. To build a small working tool that groups incoming complaints with likely duplicates, and to document its design, testing and limitations.

Research questions. RQ1: How do the three approaches differ in precision, recall and F1-score on the labelled pairs? RQ2: How do their computation costs change as the corpus grows from [n1] to [n2] complaints? RQ3: Which threshold settings give the most useful balance for the intended use?

ignouproject.com

4. Methodology

4.1 Data source and permission

Data: [name of dataset, publisher or organisation, period covered, approximate number of records, and the licence or written permission under which it may be used]. If the data come from an organisation, a permission letter from [designation] will be attached. Only text and non-identifying fields (category, date, location at [district] level) will be used.

4.2 Ethics and data protection

Complaint text can contain names, phone numbers and addresses. Before analysis these will be removed or masked using pattern rules and a manual check of a random sample. Processing will follow the obligations of the Digital Personal Data Protection Act, 2023 as applicable, and the security practices studied in MCS-078: access limited to the learner, encrypted storage and no publication of raw records.

4.3 Preprocessing

Lower-casing, removal of masked tokens and boilerplate (for example auto-generated greetings), normalisation of numbers and dates, handling of empty or very short records, and language identification to keep [English only / English and transliterated Hindi]. For MinHash, texts will be converted into word shingles of length $k = [2 \text{ or } 3]$, chosen during a pilot.

4.4 Labelled evaluation sample

A stratified random sample of [] candidate pairs will be drawn across similarity ranges and labelled by the learner as duplicate or not duplicate using written rules. A second labeller ([name, role]) will label [] pairs to check agreement, reported with Cohen's kappa.

4.5 Techniques

(a) Baseline: TF-IDF vectors with cosine similarity. (b) MinHash with [] hash functions and LSH banding (b bands of r rows), retrieving candidate pairs above a Jaccard threshold. (c) Sentence embeddings from [named pre-trained model] compared by cosine similarity, with approximate nearest-neighbour search for large corpora. Spark (MCS-071) will be used if the corpus exceeds what a single machine handles comfortably.

4.6 Evaluation

Precision, recall and F1-score for the duplicate class, precision-recall curves across thresholds, and a confusion matrix at the chosen threshold (MCS-068, Unit 12). Computation cost will be measured as wall-clock time and peak memory at corpus sizes [n1], [n2] and [n3] on stated hardware.

4.7 Tools and environment

Python [version] with pandas, scikit-learn, [MinHash library], [embedding library]; [Spark version, if used]; Git for version control; hardware [CPU, RAM, GPU if any].

ignouproject.com

5. Expected Outcomes

A documented comparison of three approaches on the same labelled data, showing for each the trade-off between accuracy and computation cost; guidance on threshold settings for this kind of short complaint text; and a clear statement of where each approach fails. No result is assumed in advance: the outcome may be that the simplest method is adequate, or that none is reliable enough without human review.

ignouproject.com

6. Deliverables

Deliverable	Description
Cleaned corpus and data dictionary	De-identified complaint text with field descriptions (kept with the learner, not published)
Labelled pair sample	[] pairs with labels, labelling rules and agreement statistic
Source code	Version-controlled notebooks and scripts for all three approaches
Grouping tool	A small application that takes a new complaint and lists likely duplicates
Project report	Full documentation of design, implementation, testing, evaluation and lessons learned

ignouproject.com

7. Proposed Work Plan

Planning suggestion only; IGNOU does not prescribe these stages or durations. Adjust with your guide.

Stage	Activities	Weeks
Data access	Permission, download, data dictionary, de-identification	1 to 3
Labelling	Pair sampling, labelling rules, second labeller	4 to 6
Implementation	Baseline, MinHash-LSH, embeddings	7 to 12
Evaluation	Metrics, thresholds, scaling tests	13 to 16
Tool and testing	Grouping tool, unit and system tests	17 to 19
Report and viva preparation	Writing, review with guide, rehearsal of the project defence	20 to 24

ignouproject.com

8. Limitations

Labels reflect the labellers' judgement and written rules; results on [dataset] may not transfer to other complaint systems or languages; pre-trained embedding models may handle code-mixed text poorly; and computation results depend on the hardware stated.

ignouproject.com

References

- Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of SEQUENCES 1997* (pp. 21-29). IEEE.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171-4186).
- Government of India. (2023). *The Digital Personal Data Protection Act, 2023* (Act No. 22 of 2023).
- IGNOU, School of Computer and Information Sciences. (2026). *Programme Guide for M.Sc. (Data Science and Analytics) (MSCDSA)*, January 2026, version 1.
- Indyk, P., and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of*

Computing (pp. 604-613).

Jurafsky, D., and Martin, J. H. *Speech and Language Processing* (3rd ed. draft).

<https://web.stanford.edu/~jurafsky/slp3/>

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2020). *Mining of Massive Datasets* (3rd ed.). Cambridge University Press.

Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 3982-3992).

Zaharia, M., Xin, R. S., Wendell, P., et al. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56-65.

[Add the recent studies you cite in Section 2.]

ignouproject.com

Enclosures to attach when submitting your own proposal

Attach exactly what the MCSP-079 project guidelines list once IGNOU publishes them (for example an approval proforma, guide details and signatures). Until then, confirm the enclosures with your Regional Centre or Learner Support Centre before posting or handing in your synopsis.

ignouproject.com