PhD Research Proposal

Data Poll: A Tool Facilitating Big Data Research

Candidate Details

Name: Mahabharathi K

Department: Computer Application

Email: mahabharathi060@gmail.com

Date of Submission: 19-09-2025

1. Introduction

The explosive growth of digital platforms and interconnected systems has resulted in a staggering amount of data generated from a variety of sources, including social media, online transactions, IoT devices, surveys, and enterprise systems. Traditional data analysis techniques often struggle to keep pace with the scale, speed, and complexity of this data. Big Data Analytics, when paired with Data Science methods, offers the essential tools needed to navigate these challenges, allowing organizations and researchers to uncover valuable insights from complex datasets.

This research puts forward the creation of DataPoll, an all-in-one tool designed to assist in the collection, processing, analysis, and visualization of big data. DataPoll aims to simplify large-scale data workflows by incorporating automated preprocessing, advanced machine learning (ML) and natural language processing (NLP) models, as well as interactive dashboards that make big data analysis more approachable and actionable.

2. Review of Literature

Big Data Analytics: Prior studies highlight the potential of big data to drive decision-making in areas such as business intelligence, healthcare, governance, and digital transformation.

Challenges of Big Data: Managing unstructured, heterogeneous, and large-scale datasets presents significant difficulties. Frameworks like Hadoop and Apache Spark provide scalable solutions but require technical expertise.

Machine Learning & NLP Applications: Techniques such as Random Forest, Support Vector Machines (SVM), XGBoost, and deep learning architectures (e.g., LSTMs) have been successfully used for classification, prediction, sentiment analysis, and topic modelling.

Research Gaps: Existing tools are often fragmented, technically complex, or focused on static datasets. Few solutions provide a fully integrated and user-friendly approach to big data analysis.

3. Objectives

The primary aim of this research is to design and implement DataPoll, a tool that simplifies big data workflows. The specific objectives are:

1. To develop scalable mechanisms for automated data collection and integration from multiple digital sources.
2. To design preprocessing modules for cleaning, filtering, and structuring large, complex datasets.
3. To apply machine learning and NLP techniques for predictive analytics, sentiment analysis, and topic modelling.

4. To create interactive dashboards and visualization tools accessible to both technical and non-technical users.
5. To validate DataPoll using real-world case studies across multiple domains (e.g., business intelligence, healthcare, governance).

4. Methodology & Research Plan

Phase 1 – Data Collection

- Collect data from APIs, web scraping, surveys, and enterprise systems.
- Incorporate both real-time and historical datasets for analysis.

Phase 2 – Data Preprocessing

- Remove duplicates, irrelevant content, and noisy data.
- Apply NLP techniques: tokenization, lemmatization, feature extraction, and sentiment tagging.
- Handle multilingual and unstructured data.

Phase 3 – Model Development

- Build predictive and analytical models using machine learning and deep learning.
- Use clustering and topic modeling to detect hidden patterns and emerging trends.

Phase 4 – Tool Development & Visualization

- Develop an intuitive, user-friendly dashboard for visualization of predictions and insights.
- Provide real-time monitoring and interactive data exploration features.

Phase 5 – Evaluation & Case Studies

- Evaluate model performance using accuracy, precision, recall, and F1-score.
- Apply DataPoll to domain-specific case studies such as customer sentiment prediction, healthcare analytics, and policy feedback monitoring.

5. Expected Outcomes

DataPoll is a handy tool designed for seamless big data collection, analysis, and visualization. It offers improved methods for preprocessing, predictive modeling, and spotting trends in real-time. This means decision-makers across various fields can access actionable insights right when they need them. Plus, it contributes to academia with publications that link big data techniques to practical research.

6. Project Summary

The DataPoll platform is set to create a flexible and cohesive space for big data research. By merging cutting-edge data science methods with easy-to-use interfaces, DataPoll aims to tackle the existing hurdles of scale, complexity, and accessibility. This initiative will enhance both academic understanding and real-world applications by facilitating effective big data analysis and providing valuable insights across various fields.