

# Sohom Ghosh

Senior Data Scientist (Experience: 9 years)

M.Tech (BITS, Pilani), PhD (Jadavpur University, Thesis Submit: May'25)

github.com/sohomghosh, Google Scholar: 7Jm4\_McAAAAJ

📍 Bengaluru, India 📞 +91-8001734384

✉ Email: sohom1ghosh@gmail.com

🏠 Web: sohomghosh.github.io

🌐 LinkedIn: sohomghosh

## SUMMARY

- **Senior Data Scientist (Individual Contributor)** with 9 years of experience pioneering innovative solutions in Artificial Intelligence (AI) to enhance digital interactions and financial solutions at scale
- **Key Achievements & Research Leadership:**
  - Delivered **20+** end-to-end data science **projects**, including **9** solutions deployed in **production** impacting millions of users, Patented solution for NLP-driven call transcripts analytics
  - Earned **10 awards** & completed **11 advanced certifications** (MOOCs) in AI, Natural Language Processing (NLP), & Machine Learning (ML)
  - Published **31 peer-reviewed papers** at top venues: **TheWebConf (WWW), CIKM, COLING, LREC, IEEE Big Data, CODS-COMAD, etc.** (📄Google Scholar: **200+** citations, **h-index: 7**)
  - Co-authored **2 technical books** on NLP, Got **1 US patent** granted for NLP-driven call analytics innovation
  - Technical Focus: Applications of **Large Language Models (LLMs), NLP, Generative AI, & ML**
  - Key Skills: **Large Language Models | Multi-modal AI | Multi-lingual NLP | Retrieval-Augmented Generation (RAG) | MLOps (basics) | Financial NLP | PyTorch | AWS | AI Agents**

## WORK EXPERIENCE

### • Fidelity Investments

Jun 2019 - Present

Senior Analyst → Data Scientist → Senior Data Scientist

Bengaluru, India

#### \* **Generative AI for analysing Call and Live Chat Transcripts** | 📄Patented | In Production

- Designed and managed an annotation pipeline using **Appen**, overseeing a team of 98 annotators to ensure **high-quality labelled datasets** for model training and evaluation
- Fine-tuned advanced NLP models, including **Bi-LSTM, T5, BART, and LLaMA**, for **multi-theme extraction and summarization** from call transcripts. Optimized model performance to accurately identify and condense key insights, enhancing decision-making and workflow efficiency. Applied **Hierarchical Clustering** recursively to group similar themes and identify patterns within unstructured data. Productionized NLP models using MLOps frameworks. Led end-to-end LLM pipelines (fine-tuning, evaluation, deployment).
- **Applications:** i) **Automated short note generation** for call transcripts to streamline communication and improve accessibility ii) Analysing **root causes of high call volumes** to identify trends and optimize operational efficiency iii) **Featurization of textual interaction data** to enable advanced analytics and actionable insights
- **Impact:** i) Identified and resolved login-related issues, resulting in a **10% increase in customer satisfaction scores** ii) Updated workflows for key processes, driving an **18-point increase in Net Promoter Score (NPS)** and a **27% improvement in Customer Ease Score** iii) Developed a comprehensive solution to address the student debt crisis, leading to a **60% increase in enrolment**, **59,000 yearly payments** processed, and **\$200M+** disbursed to participants.

#### \* **Predictive Customer Interaction System** | 📄In Production

- Experimented with **pattern mining algorithms, LLM, Learning-to-Rank** frameworks & engineered real-time solutions to predict customer's next-step actions during an interaction with chatbot. Identifying content gap for aiding content creation (**32.5% Click-Through Rate**, Increased Customer Satisfaction Score by **6%**, Estimated savings **\$300M** by avoiding calls)
- Developed **statistical & linguistic metrics** to measure effectiveness of automated chat sessions

#### \* **Investor Behaviour Analytics, Enterprise Search & Knowledge Solutions**

- Developed a **LightGBM-based predictive model** leveraging user profiles and web activity data to **forecast equity sell-offs**. Enabled proactive decision-making by accurately identifying patterns in user behaviour
- Experimented with **Retrieval-Augmented Generation (RAG)** to enable accurate and context-aware responses to **business queries from call transcripts**. Explored and implemented **advanced chunking and denoising strategies** to optimize data preprocessing. Leveraged frameworks such as LangChain, LlamaIndex, FAISS, and Chroma DB to enhance retrieval efficiency and model performance
- Conducted **trend analysis and categorization of 3 million investor search queries** to prioritize content creation, contributing to a nominated paper for the Gartner Eye on Innovation Awards for Financial Services 2023 (one of three nominations submitted by Fidelity Investments) [📄paper]
- Developed a **RoBERTa-based direct answer algorithm** to improve the search experience, enabling users to find relevant information faster and more accurately [📄paper]

**Technical Stack:** Python, PyTorch, AWS, Snowflake, SQL; **Won 7 Awards: 7, Deployed 4 production ML models, Published 3 research papers, Keywords:** LLMs, RAG, Clustering, Predictive Analytics, NLP

## • Times Internet

Jan 2017 - Jun 2019

Data Scientist

Noida, India

- \* Designed and implemented a **Word2Vec-based skill recommendation system** for TimesJobs to enhance search experience of candidates
- \* Engineered a scalable **XGBoost**-based predictive model using **PySpark** for email targeting, deployed on a **Hadoop** cluster (achieving a **15% increase** in email open rates)
- \* Delivered actionable insights that informed strategic decisions by analysing **interest graphs and behavioural patterns** of **450M+ monthly visitors** across 39+ digital properties including Gaana, Times of India, Economic Times, MX Player, and CricBuzz
- \* Spearheaded multiple analytics initiatives:
  - **Sales Analytics**: Upsell/cross-sell strategies and **XGBoost-based** churn modelling
  - **Digital Product Analytics, B2B Cross-walk Analytics**
  - **Fraud Analytics**: Detecting fraudulent activities in affiliate marketing (**preventing \$17,000+ in losses**)




## • Fn MathLogic

Jul 2016 - Jan 2017



Analyst






Gurugram, India

## EDUCATION

- **PhD in Engineering, Topic: Financial Natural Language Processing** ( summary poster video) 2022-25 (tentative)  
Jadavpur University, Kolkata, India
- **Master of Technology in Software Systems (Data Analytics)** 2017-19  
BITS, Pilani, India (WILP)
- **Bachelor of Technology in Computer Science & Engineering** 2012-16  
Heritage Institute of Technology (Maulana Abul Kalam Azad University of Technology), Kolkata, India

## SELECTED PERSONAL PROJECTS & PUBLICATIONS

(For more: GitHub, Google Scholar)

- **Generator-Guided Crowd Reaction Assessment** (In TheWebConf (WWW-2024), paper)  
Finetuned FLANG-Roberta with responses from LLMs (Claude, ChatGPT, Flan-UL2) using cross-encoders for tweet engagement prediction
- **IndicFinNLP: Financial Natural Language Processing for Indian Languages** (In LREC-COLING 2024, paper)  
Created datasets in Hindi, Bengali, & Telugu for analysing argumentative posts, assessing sustainability, detecting exaggerated numerals & ESG themes
- **Experimenting with Multi-modal Information to Predict Success of Indian IPOs** (data, code, pre-print)  
Proposed a RAG, NLP, & ML-based approach to predict IPO success by analysing factors like IPO prospectus details, macro-economic conditions, and market data. Created two new datasets for Indian IPOs and explored multi-modalities (text, images, numbers, & categorical features) to estimate IPO performance metrics such as direction and underpricing.

## TECHNICAL SKILLS AND INTERESTS

**Languages & Libraries** : Core (Python, PyTorch, SQL & Cloud (AWS)), ML (Scikit-learn, Pandas, Numpy, XGBoost, LightGBM), NLP (NLTK, SpaCy, Transformers, LangChain), UI (Gradio, Streamlit), Spark

**Algorithms & Concepts**: Regression (Linear/Logistic), Decision Trees, Random Forest, Gradient Boosting Machine, Clustering, PCA, Neural Networks, Deep Learning, Large Language Models, Prompt Engineering, RAG, Fine-tuning

**Others**: LaTeX, MS Office (Word, Excel, PowerPoint), Confluence, Git, Jira, Kanban, Mural

**Expertise**: Natural Language Processing / Understanding / Generation

**Domains**: FinTech (Financial Services + Technology), Consumer Internet based Products & Customer Analytics

**Relevant Courses**: **Technical - Natural Language Processing Specialization** (Score >90%), **Neural Networks and Deep Learning, Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization**

(Score>90%), **Prompt Engineering, Large Language Models, Cloud - AWS, Machine Learning Engineering, etc., AI Agents** (LangGraph, crewAI); **Soft skills - Learnship Business English Level 10, Creating Effective Presentations, Creative Thinking; Research - Introduction to Research, NPTEL, Score: 87%, Rank: Top 1%, Domain - Mutual Funds, Stocks etc.**

**Areas of Interest**: Applied NLP, GenAI, ML

**Soft Skills**: Critical thinking, Communication, Problem Solving, Self-learning, Resilience, Emotional intelligence

## ACHIEVEMENTS & EXTRACURRICULAR ACTIVITIES

**Awards**:CODS-COMAD-2023 YRS Honourable Mention; Travel Grants- ARCS-2025, PIC-2025, CIKM-2024, CODS-COMAD-2024; Fidelity Investments- Kudos [2024], Eureka Innovation Enablers [2023], On the Spot [2023], Excellence in Action (2X) & You've Earned It (2X Individual, 2XTeam), Shout Out [2020, 2021, 2023], Patent Award [2021, 2024]; **Times Internet- Rock Star** [2018], **Hackathons**: Kaggle: Datasets Expert & 7 Bronze, TechGig: CodeGladiators-2018 Finalists, **Analytics Vidhya**: 4 times in top 25  
**Extracurricular Activities**: Playing Harmonica (ENERGIZE: Rank-2@2024, Finalist@2025), Martial Arts, Completed 11 Treks (received IndiaHikes Green Getter & Trekker for Life Awards)