

# Sohom Ghosh

SENIOR DATA SCIENTIST (EXPERIENCE: 8.5+ YEARS) · M.TECH (BITS, PILANI) · PHD (PURSUING, JADAVPUR UNIVERSITY)

☎ (+91) 8001734384 | ✉ sohom1ghosh@gmail.com | 🏠 sohomghosh.github.io | 📄 sohmghosh | 🌐 sohomghosh | 🎓 Google-Scholar

## Summary

- **Senior Data Scientist (Individual Contributor)** responsible for improving the digital lives and financial well-being of millions of users
- Projects: **20 (delivered)** and **8 (in production)**; Completed **11 MOOCs/Certifications**; Received **10 Awards**
- Research Interests: **Applications of Generative Artificial Intelligence, Large Language Models (LLMs) & Natural Language Processing**
- Co-authored **2 books**, Published **31 research papers**, Filed **2 US patents**, Received **188 citations** on [Google Scholar](#)  
(Venues: **TheWebConf (WWW)**, **CIKM**, **COLING**, **LREC**, **IEEE Big Data**, **CODS-COMAD** etc.)

## Work Experience

### Fidelity Investments

Bengaluru, India

SENIOR ANALYST → DATA SCIENTIST → SENIOR DATA SCIENTIST

Jun. 2019 - Present

- Developed and implemented content-based, collaborative filtering, and pattern mining algorithms. Used Large Language Models & Learning to Rank based approaches to predict the next interaction of a customer during a chat session.
- Experimented with Large Language Models (LLM) (**T5**, **Llama**). Created **Bi-LSTM**, **T5**, & **BART** based models for **summarizing** & extracting multiple themes from call transcripts. Used **Agglomerative Hierarchical Clustering** recursively to group similar themes. Designed the annotation job using **Appen** and administered a team of **98 annotators**. **Applications**: i) Automated short note generation for calls, ii) Comprehending the reasons behind high call volumes, iii) Featurization of textual interaction data. **Impact**: i) Issues related to logging in were identified & resolved leading to a **10% increase** in customer satisfaction score ii) Workflows for different processes were updated leading to an **increase** in Net Promoter Score and Customer Ease Score by **18 points** & **27%** respectively iii) Competent solution to address the student debt crisis was created which lead to **60% increase** in enrollment, **59K** yearly payments & **\$200M+** being paid to the participants [paper](#) [news](#) [patent](#)
- Experimented with Retrieval Augmented Generation (**RAG**) to answer business queries from call transcripts. Worked on various chunking and denoising strategies. Explored several frameworks like LangChain, LlamaIndex, FAISS, Chroma DB, etc.
- **Trend analysis & categorization** of **3 million** search queries of investors for prioritizing **content creation** [paper](#) (One of the three nominations sent by Fidelity Investments for Gartner Eye on Innovation Awards for Financial Services - 2023)
- Developed a **RoBERTa**-based **direct answer** algorithm to enhance the search experience of users. [paper](#)
- Built a **Light-GBM**-based model from users' profiles & web-activities to predict equity sell-off
- **Tools**: Python, PyTorch, Amazon Web Services (AWS), Snowflake, SQL, MS Office, Git & Jira
- **Awards**: 7; **Models in production**: 4; **Keywords**: **Large Language Models, Summarization, Question Answering, Clustering, Prediction**

### Times Internet

Noida, India

DATA SCIENTIST

Jan. 2017 - Jun. 2019

- Developed a **Word2Vec**-based **skill recommendation system** for TimesJobs
- Built an **XG-Boost**-based predictive model using **PySpark** to target emails & deployed it on a **Hadoop** cluster thereby increasing the open rate
- Presented **insights** for impacting business decisions by analyzing interest graphs & behaviours of 450+ million monthly visitors across 39+ digital products (Gaana, Times of India, Economic Times, MX Player, CricBuzz etc.)
- Other projects: **Sales Analytics** (Upsell, Cross-sell, **XG-Boost** based Churn Model development), **Digital Product Analytics**, **B2B Cross-walk Analytics**, **Fraud Analytics** in affiliate marketing (**saved \$17,000**), Career Graph, Data Engineering
- **Models in production**: 3; Pursued **work integrated M.Tech**

### Fn MathLogic

Gurugram, India

ANALYST

Jul 2016 - Jan 2017

### Selected Personal Projects

More projects in [GitHub](#).

- **Crowd Reaction Analysis** using **Generative Large Language Models** [WWW-2024 pre-print](#)  
→ Finetuned **FLANG-Roberta** with responses from **LLMs** (Claude, ChatGPT, Flan-UL2) using cross-encoders to predict which version of a text will receive more engagement
- **Hypernym Detection** in Financial Texts [paper](#)  
→ Finetuned **FinBERT** embeddings using **sentence-transformers** to detect generic forms of financial keywords
- **Improved Investing**: Evaluating social media posts by Executives [paper](#), Assessing profitability from social media posts [paper](#), Claim & Exaggerated Numeral Detection ([paper](#), [pre-print](#), [demo](#)), Argument Analysis [paper](#)  
→ Proposed several pre-trained language model based solutions to counter finance related misinformation in social media  
→ Benchmarked performances of **LLMs** (Dolly, MPT, Flan T5) with **BERT-SEC**, **FLANG-RoBERTa**, etc.
- Financial **Readability Assessment** Dataset & **Text Simplification** [code & data](#) [paper](#) [demo](#)  
→ Proposed a dataset (FinRAD) & an architecture to assess the readability of financial texts and benchmarked it with standard rule-based readability scores (Flesch, SMOG, etc.)
- Financial Language **Understandability Enhancement** Toolkit [paper](#) [demo](#)  
→ Developed a toolkit using Gradio capable of summarizing, detecting hypernyms & in-claim numerals, extracting sentiments, assessing readability & sustainability, extracting ESG & Forward looking statements from financial texts
- Financial Natural Language Processing for **Indian Languages** [code & demo](#) [data](#) [paper-1](#) [paper-2](#)  
→ Created datasets in Hindi, Bengali, & Telugu for analysing argumentative posts, assessing sustainability, detecting exaggerated numerals & ESG themes
- Data driven approach for predicting success of Indian IPOs  
→ Used RAG, LLM, & Machine Learning

## Technical Skills

- Tools & Technologies: **Python, SQL & Cloud (AWS)**
- Libraries & Frameworks: **PyTorch, Scikit-learn, Pandas, Numpy, XG-Boost, LightGBM, NLTK, SpaCy, Gradio, Spark, Hadoop, Transformers**
- Algorithms & Concepts: **Regression (Linear/Logistic), Decision Trees, Random Forest, Gradient Boosting Machine, XGBoost, Clustering, PCA, Neural Networks, Deep Learning, Large Language Models, Prompt Engineering, Retrieval Augmented Generation**
- 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**

## Education

### Jadavpur University

**PHD [pursuing] IN ENGINEERING** (TOPIC: **FINANCIAL NATURAL LANGUAGE PROCESSING**) [SUMMARY](#) [POSTER](#) [VIDEO](#)

Kolkata, India

2025 (tentative)

### BITS, Pilani (WILP)

**M.TECH IN SOFTWARE SYSTEMS** (SPECIALIZATION: **DATA ANALYTICS**)

Pilani, India

2019

### Heritage Institute of Technology (Maulana Abul Kalam Azad University of Technology)

**B.TECH IN COMPUTER SCIENCE & ENGINEERING**

Kolkata, India

2016

## Selected Publications, Achievements, Extracurricular Activities

### NOTABLE PUBLICATIONS

† means co-first authors

- **S Ghosh**, et al., "Generator-Guided Crowd Reaction Assessment", in **TheWebConf (WWW-2024)**, Singapore [paper](#)
- **S Ghosh**, "Demystifying Financial Texts Using Natural Language Processing", in **CIKM 2024**, Boise, USA [pre-print](#)
- **S Ghosh**, et al., "IndicFinNLP: Financial Natural Language Processing for Indian Languages", in **LREC-COLING 2024**, Torino, Italy [paper](#)
- **S Ghosh**, et al., "FLUEnT: Financial Language Understandability Enhancement Toolkit", in **CODS-COMAD 2023**, Mumbai, India [paper](#) [demo](#)
- **S Ghosh**, A Chopra, S K Naskar, "Learning to Rank Hypernyms of Financial Terms Using Semantic Textual Similarity", in **Springer Nature Computer Science** journal, 2023 [paper](#) [code](#)
- A Singh, C Bhatia, **S Ghosh**, "The Effect of Tweets on the Traded Volume of Crypto-Coins", in **IEEE Big Data 2023**, Sorrento, Italy
- **S Ghosh**, et al., "FinRAD: Financial Readability Assessment Dataset - 13,000+ Definitions of Financial Terms for Measuring Readability", in **FNP** workshop of **LREC-2022**, Marseille, France [code & data](#) [paper](#)
- A Chopra, D Bal, **S Ghosh**, "Automated Analysis of Customer Interaction Text to Generate Customer Intent Information and Hierarchy of Customer Issues", **US Patent #**: 12033162 [link](#)
- **S Ghosh**†, A Chopra†, "Using Transformer Based Ensemble Learning to Classify Scientific Articles", in Trends and Applications in Knowledge Discovery and Data Mining, **PAKDD-2021**, Delhi, India [paper](#) [code](#) [video](#)
- R Chopra, **S Ghosh**, et al. "The Natural Language Processing Workshop" (**Book**), Packt Pub., [2020] [link](#)
- **S Ghosh**, D Gunning, "Natural Language Processing Fundamentals" (**Book**), Packt Pub., [2019] [link](#) [code](#)

### ACHIEVEMENTS & EXTRACURRICULAR ACTIVITIES

- Awards: **Kaggle**-Datasets Expert [2024]; Travel Grants- **ARCS-2025**, **PIC-2025**, **IndoML-2024**, **CIKM-2024**, **CODS-COMAD-2024**; **CODS-COMAD YRS Honourable Mention** [2023]; **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 Filing Award** [2021]; **Times Internet** (TBS)- **Rock Star Award** [2018]
- Hackathons: **Kaggle**: 7 Bronze, **TechGig**: CodeGladiators-2018 Finalists, **Analytics Vidhya**: 4 times in top 25
- 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, Computer Vision, etc.; **Soft skills** - Learnship Business English Level 10, Creating Effective Presentations, Creative Thinking; **Research** - Introduction to Research, NPTEL, Score: 87%, **Rank: Top 1%** (Elite + Silver) (Jan 2021) [link](#), **Domain** - Mutual Funds, Stocks etc.
- Extracurricular Activities: Playing **Harmonica** (**ENERGIZE-2024** - Rank-2), Completed 11 **Treks** (received IndiaHikes Green Getter & Trekker for Life Awards)