

# Sabur Butt

✉ saboor.butt007@gmail.com

📧 saboor.butt2

🌐 <https://saburbutt.com>

github.io/ <https://linkedin.com/in/saburb>

📍 Monterrey, Nuevo Leon

📞 +52 (1) 5562298639



## About Me

I am a Natural language processing expert with 4 years of experience working on a variety of data science projects, including developing machine learning models, performing statistical analysis, python development and visualizing data. I am passionate about using data to solve complex problems and drive meaningful impact. I have the skills and knowledge to extract insights and patterns from large datasets, and to communicate those insights in a clear and actionable way.



## Education

- Feb 2021 – Oct 2023 📖 **Ph.D. Computer Science** Instituto Politecnico Nacional, Mexico City, Mexico  
Thesis title: *Automatic Personality and Behavior Detection in Text*  
Advisor: *Dr. Grigori Sidorov.*
- Jan 2019 – Jan 2021 📖 **M.Sc. Computer Science** Instituto Politecnico Nacional, Mexico City, Mexico  
Thesis title: *Question Answering in Open Domain*  
Advisor: *Dr. Grigori Sidorov.*
- Sept 2013 – Dec 2017 📖 **B.Sc. Computer Science** Forman Christian College, Lahore, Pakistan  
Thesis title: *Brain tumor segmentation and classification*






## Employment History

- Nov 2023 – Ongoing 📖 **Postdoctoral Researcher** Tecnológico de Monterrey, Mexico
  - I am currently spearheading interdisciplinary projects that focus on the intersection of Natural Language Processing and Education.
  - Technologies used: [LangChain](#), [Streamlit](#), [Llama-2](#), [Keras](#), [TensorFlow](#)
- Aug 2022 – May 2023 📖 **Research Stay.** Tecnológico de Monterrey, Mexico
  - Led a project on Natural Language Processing for Student Evaluation of Teachings, resulting in two conference papers.
  - Technologies used: [Python](#), [GPT-3](#), [BERT](#), [scikit-learn](#), [Keras](#), [TensorFlow](#)
- July 2021 – Dec 2021 📖 **Academic Collaboration.** University of Tartu, Estonia
  - Produced a journal paper on rumour detection in text
  - Technologies used: [Python](#), [Tableau](#), [LIWC](#), [SenticNeT](#), [Emotion Detection](#), [SHAP](#), [Random Forest](#), [scikit-learn](#), [Keras](#), [TensorFlow](#)
- July 2020 – Dec 2020 📖 **Academic Collaboration.** Chang Gung University, Taoyuan, Taiwan
  - Created a novel dataset for multi-label emotion detection in Urdu Language. Produced a journal paper with machine / deep learning baselines
  - Technologies used: [Python](#), [Random forest \(RF\)](#), [Decision tree \(J48\)](#), [Sequential minimal optimization \(SMO\)](#), [AdaBoostM1](#), and [Bagging](#), [Convolutional Neural Networks \(1D-CNN\)](#), [Long short-term memory \(LSTM\)](#), [BERT](#), [scikit-learn](#), [Keras](#), [TensorFlow](#)

## Employment History (continued)



- June 2017 – Jan 2019  **Research Associate.** Intelligent Machines Lab, ITU, Lahore, Pakistan  
- Lead research in Human Robot Interaction and Artificial Intelligence. [Link](#)  
- Technologies used: [Visual Studio](#), [C#](#), [Adobe Character Builder](#), [Mixamo](#), [SQL](#), [MySQL](#), [Kinect](#), [Object Detection](#)
- Feb 2016- Feb 2017  **Web Developer.** Acumen Solutionz, Lahore, Pakistan  
- Responsible for creating user-friendly designs and developing websites. Made several websites for local businesses. From creating layouts to functions according to clients specifications  
- Technologies used: [WordPress](#), [Woocommerce](#), [JavaScript](#), [HTML](#), [CSS](#), [SQL](#)

## Skills






- Languages  English (C1), Urdu (Native), Punjabi (Native), Spanish (A2).
- Coding  Python, Java, MATLAB, C#, SQL, XML,  $\LaTeX$ , ...
- Databases  MySQL, SQLite.
- Web Dev  HTML, CSS, JavaScript, Apache Web Server.
- Technologies  PyTorch, scikit-learn, Keras, NLTK, spaCy, CoreNLP, Fastai, HuggingFace, git, Tableau, LangChain, Streamlit.

## Professional Achievements and Contributions


### Academic Grants

- 2021  **Principal Investigator (PI):** Two-year technical development and innovation grant for the project “VOISELL - System for the accessibility of digital markets for people with loss of vision” [Link](#)
-  **Co-Investigator:** Two-year technical development and innovation grant for the project “Automatic vocabulary creation system to facilitate learning lexicon of humans and machines” [Link](#)










### Awards and Scholarships

- 2016  **Travel grant and living allowance,** By Forman Christian College, Lahore, Pakistan, for representing the institute in United Asian Debating Championship, Bangkok, Thailand.
- 2017  **Roll of Honor,** By Forman Christian College, Lahore, Pakistan, for outstanding services in representing the institute in parliamentary style debates in more than 30 national competitions.
-  **Services Award and Core Values Award,** By Forman Christian College, Lahore, Pakistan for leading national debating society roles (Finance Director and General Secretary).
- 2019  **M.Sc and Ph.D. scholarship of the CONACYT,** Government of Mexico 2019-2023, for studying in the Masters and Ph.D. program certified by the CONACYT (Ministry of Science)
- 2021  **Honorable Mention at Facebook Hackathon,** For Project [Voisell](#) facilitating visually impaired to buy and sell things online using NLU and Wit.AI



### Leadership Experience

- 2017  **Lead Robotics at 7th and 8th Robotics Expo,** Mega Robotic Event, Lahore, Pakistan. Took this initiative to empower the young kids to solve locally relevant problems of Pakistan with custom-built robots. Lead all the phases of robot development such as proposals, user requirements, design, development, documentation, webpage and delivery. [Link](#)

## Professional Achievements and Contributions (continued)

- 2020  **Task Committee member at CICLing**, UrduFake track @FIRE 2020, Mexico. Managed the website, task entries and evaluation reports by all participants. [Link](#)
- 2020  **Associated member of AMPLN**, Mexican Association of Natural Language Processing. [Link](#)
- 2021  **Organizer at CICLing** UrduFake news detection track @FIRE. [Link](#)
- 2021  **Organizer at CICLing** Urdu threat and abuse track @FIRE 2021 co-hosted with ODS SoC 2021. [Link](#)
- 2022  **Organizer at CICLing** Organizer at CICLing Multi-label emotion and threat language detection task (Emothreat) @FIRE 2022. [Link](#)
- 2022  **Organizer at CoLI-Kanglish** Word Level Language Identification in Code-mixed Kannada-English Texts @ICON 2022. [Link](#)
- 2023  **Organizer at CoLI-Tunglish** Word-level Language Identification in Code-mixed Tulu Texts @FIRE 2023. [Link](#)
- 2024  **Organizer of HOPE at IberLEF** Approaching Hope Speech Detection in Social Media from Two Perspectives, for Equality, Diversity and Inclusion and as Expectations @SEPLN. [Link](#)
- 2024  **Organizer at CoLI-Dravidian** Word-level Code-Mixed Language Identification in Dravidian Languages @FIRE 2024. [Link](#)


## Roles for Journals and Conferences

- 2021-23  Reviewer for LKE'2021: 8th International Symposium on Language & Knowledge Engineering, IEEE Access, Research in Computing Science, MICAI: Mexican International Conference on Artificial Intelligence 2021-2023, FIRE: Forum for Information Retrieval Evaluation, 2021-2023, PLOS One
- 2023  Editorial board member of PriMera Scientific Engineering [Link](#)

## Invited Talks and Seminars

- May 2023  Seminar presented on “Emerging Technologies and Concepts for Large Language Models (LLM)” at IFE Living Lab and Data Hub, Tecnológico de Monterrey, Mexico

## Advising and Mentoring

- 2024-2025  Diana Patricia Madera Espindola (Masters) “Large Language Models for identifying skills in resumes and job postings” at Tecnológico de Monterrey, Mexico

## Research Publications

- 1 Balouchzahi, F., **Butt, S.**, Sidorov, G., & Gelbukh, A. (2023). Reddit: Regret detection and domain identification from text. *Expert Systems with Applications*, 225, 120099.
- 2 **Butt, S.**, Mejia-Almada, P., Alvarado-Uribe, J., Ceballos, H. G., Sidorov, G., & Gelbukh, A. (2023). MF-SET: A multitask learning framework for student evaluation of teaching. *Proceedings of the Future Technologies Conference*, 254–270.
- 3 Gallardo, K., **Butt, S.**, & Ceballos, H. (2023). Improvement of teaching competencies training in higher education faculty based on student evaluations of teaching and ai systems. *International Conference in Information Technology and Education*, 555–563.

- 4 Hegde, A., Balouchzahi, F., Coelho, S., HL, S., Nayel, H. A., & **Butt, S.** (2023). CoLI@ FIRE2023: Findings of Word-level Language Identification in Code-mixed Tulu Text. *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, 25–26.
- 5 Hegde, A., Balouchzahi, F., Coelho, S., Shashirekha, H., Nayel, H. A., & **Butt, S.** (2023). Overview of CoLI-Tunglish: Word-level Language Identification in Code-mixed Tulu Text at FIRE 2023. *FIRE (Working Notes)*, 179–190.
- 6 Sidorov, G., Balouchzahi, F., **Butt, S.**, & Gelbukh, A. (2023). Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets. *Applied Sciences*, 13(6), 3983.
- 7 Amjad, M., **Butt, S.**, Zhila, A., Sidorov, G., Chanona-Hernandez, L., & Gelbukh, A. (2022). Survey of fake news datasets and detection methods in european and asian languages. *Acta Polytechnica Hungarica*, 19(10), 185–204.
- 8 Ashraf, N., Khan, L., **Butt, S.**, Chang, H.-T., Sidorov, G., & Gelbukh, A. (2022). Multi-label emotion classification of urdu tweets. *PeerJ Computer Science*, 8, e896.
- 9 Ashraf, N., Rafiq, A., **Butt, S.**, Shehzad, H. M. F., Sidorov, G., & Gelbukh, A. (2022). Youtube based religious hate speech and extremism detection dataset with machine learning baselines. *Journal of Intelligent & Fuzzy Systems*, 42(5), 4769–4777.
- 10 Balouchzahi, F., **Butt, S.**, Hegde, A., Ashraf, N., Shashirekha, H., Sidorov, G., & Gelbukh, A. (2022). Overview of coli-kanglish: Word level language identification in code-mixed kannada-english texts at icon 2022. *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, 38–45.
- 11 Balouchzahi, F., **Butt, S.**, Sidorov, G., & Gelbukh, A. (2022). CIC@LT-EDI-ACL2022: Are transformers the only hope? hope speech detection for spanish and english comments. *LTEDI 2022*, 206.
- 12 **Butt, S.**, Amjad, M., Balouchzahi, F., Ashraf, N., Sharma, R., Sidorov, G., & Gelbukh, A. (2022a). Emothreat@ fire2022: Shared track on emotions and threat detection in urdu. *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, 1–3.
- 13 **Butt, S.**, Amjad, M., Balouchzahi, F., Ashraf, N., Sharma, R., Sidorov, G., & Gelbukh, A. (2022b). Overview of emothreat: Emotions and threat detection in urdu at fire 2022. *Proceedings of the CEUR Workshop Proceedings, Chennai, India*, 22–24.
- 14 **Butt, S.**, Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2022). Cic@ pan: Simplifying irony profiling using twitter data. *CEUR Workshop Proceedings*, 3180, 2402–2410.
- 15 **Butt, S.**, Sharma, S., Sharma, R., Sidorov, G., & Gelbukh, A. (2022). What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses. *Computers in Human Behavior*, 107345.
- 16 Thang Ta, H., **Butt, S.**, Jason, A., Sidorov, G., & Gelbukh, A. (2022). The combination of BERT and data oversampling for relation set prediction. *20th International Semantic Web Conference*.
- 17 **Butt, S.**, Ashraf, N., Fahim, H., Sidorov, G., & Gelbukh, A. (2021). Transformer-based extractive social media question answering on TweetQA. *Computación y Sistemas*, 25(1).
- 18 Amjad, M., Alisa, Z., Sidorov, G., Andrey, L., **Butt, S.**, Hamza Imam, A., Oxana, V., & Gelbukh, A. (2021). Overview of abusive and threatening language detection in urdu at fire 2021. *CEUR Workshop Proceedings*.
- 19 Amjad, M., **Butt, S.**, Amjad, H. I., Zhila, A., Sidorov, G., & Gelbukh, A. (2021). Urdufake@fire2021: Shared track on fake news identification in urdu. *Forum for Information Retrieval Evaluation*, 19–21. <https://doi.org/10.1145/3503162.3505240>
- 20 Amjad, M., **Butt, S.**, Hamza Imam, A., Alisa, Z., Sidorov, G., & Gelbukh, A. (2021). Overview of the shared task on fake news detection in urdu at fire. *CEUR Workshop Proceedings*.

- 21 Amjad, M., Zhila, A., Sidorov, G., Labunets, A., **Butt, S.**, Amjad, H. I., Vitman, O., & Gelbukh, A. (2021). Urdu threat@fire2021: Shared track on abusive threat identification in urdu. *Forum for Information Retrieval Evaluation*, 9–11. <https://doi.org/10.1145/3503162.3505241>
- 22 Ashraf, N., **Butt, S.**, Sidorov, G., & Gelbukh, A. (2021). CIC at checkthat! 2021: Fake news detection using machine learning and data augmentation. *CLEF, 2021–Conference and Labs of the Evaluation Forum*.
- 23 **Butt, S.**, Ashraf, N., Sidorov, G., & Gelbukh, A. (2021). Sexism identification using bert and data augmentation–exist2021. *International Conference of the Spanish Society for Natural Language Processing SEPLN*.
- 24 Malik, Z. H., **Butt, S.**, & Sajid, H. (2019). Quality scale for rubric based evaluation in capstone project of computer science. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent computing* (pp. 219–233). Springer International Publishing.
- 25 Rehmani, T., **Butt, S.**, Baig, I.-R., Malik, M. Z., & Ali, M. (2018). Designing robot receptionist for overcoming poor infrastructure, low literacy and low rate of female interaction. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 211–212.