# CURRICULUM VITAE

## PERSONAL INFORMATION

| | |
|---|---|
| Name | Mohammed Zuhair Naji Al-Taie |
| Birth | 1977 |
| Address | Al-Khadrah neighborhood / Baghdad / Iraq |
| E-mail | mza004@live.aul.edu.lb |
| Hand Phone | 07802969686 |
| Gender | Male |
| Marital Status | Married and have 2 children (12 and 8 years) |

## FIELDS OF INTEREST

**Fields of interest:**

- Data Science,
- Social Networks,
- Big Data,
- Machine Learning,
- Data Mining,
- Technology Networks,
- Web Technologies.

**Skills**:

- Data science techniques,
- Big Data technologies including Apache Hadoop, Apache Spark, Apache Hive, and Apache Pig,
- Database administration,
- Analysis tools including R and SPSS,
- Programming languages including Python, C++, and Java,
- Preparing project proposals and reports,
- University teacher,
- Team leader,
- Excellent communication skills in English, both verbal and written,
- Presentation skills,
- Researcher and author,
- Business oriented,
- Volunteer, and
- Blogger.

## EDUCATIONAL QUALIFICATIONS

| | |
|---|---|
| 2018 | Ph.D in computer science: Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, Johor. |
| 2012 | M.S in Computer Science and Communication: Arts Science & Technology University in Lebanon (AUL). |
| 2006 | BSc in Computer Science: Al-Mustansiriya University, College of Science, Computer Sciences Department. |
| 1996 | Secondary School Grade (Baghdad College Secondary School for boys) |

| WORK EXPERIENCE (2000 – 2018) | | |
|---|---|---|
| Aug 2019 – up to now | | Senior data scientist/Analytics group/Business Intelligence department/Zain Iraq. |
| 1 March 2019 – July 2019 | | University teacher at the computer engineering department. |
| May 2018 – Nov 2018 | | Data scientist at Axiata Analytics Center (AAC), Axiata Group Berhad (Axiata Group of Companies), based in Kuala Lumpur – Malaysia (details of projects implemented are shown in the next table). |
| April 2014 – May 2018 | | During my PhD study, and as a part-time job, I was involved in a number of projects in cooperation with UTM Big Data Center. The goal of establishing UTM Big Data Center was to: intersect the core of Big Data Analytics into big data science, big data computing and big data engineering, as well as to provide the right information at the right time, which enables data analysts, data scientist, data engineers or managers to make informed business decisions (details of projects implemented are shown in the next table). |
| Oct 2012 – Jan 2014 | | Teacher at Al-Salam University College (Computer Science Department) in Baghdad, Iraq, teaching courses in artificial intelligence, software engineering, computer networks, and progamming. |
| Mar 2012 – Oct 2012 | | Teacher at Al-Khalid private institute in Baghdad, Iraq, for training. My responsibilities included giving lectures on artificial intelligence, Web design and development and computer networking. |
| Feb 2010 – Feb 2012 | | Doing my Master of Science degree in Lebanon. |
| Dec 2009 – Feb 2010 | | Waiting to start my Master study |
| July 2008 – Dec 2009 | | System developer and database administrator at 'Taj Al-Khairat Group' for General Trading & Contracting.<br>My typical responsibilities as a systems developer were discussing the client's requirements and proposing solutions, using development tools to write and test code to develop these solutions, writing documentation and operating manuals, and testing and modifying systems to ensure that they operate reliably.<br>My typical responsibilities as a database administrator where distributed over ongoing maintenance of production databases; planning, design, and development of new database applications; and management of the organization's data and metadata. |
| Oct 2007 – July 2008 | | IT specialist and team leader at 'Advanced Technologies Company' for internet services.<br>My duties as an IT specialist in this company included offering assistance to computer system users, keeping current systems updated and running smoothly, and overseeing the organizations data, including client information and internal information.<br>One of the main projects that we worked on was to provide 15 government locations across Iraq with PCs, printers, scanners, hubs, connections and other requirements. The project also included connecting these locations to the local internet provider. |
| Nov 2006 – Oct 2007 | | IT specialist and team leader at 'Al-Baghdadya Company' for Information Technology services.<br>My typical responsibilities in this company involved network management, software development, technical support to the organization's employees and train non-technical workers on the business's information systems. The job also included designing systems and assessing the effectiveness of technology resources already in use or new systems that are being implemented.<br>One of the projects that we completed was building a VSAT network that connects Iraq and Kuwait. The aim of the project was to connect several banks in Iraq to a server location in Kuwait through a VSAT network. With the technical support from ITS and Streamlink companies in Kuwait, this project helped storing customer transactions in a safe and maintainable environment. Besides construction works, the work also included the provision and installation of routers, switches, server rooms, UPS rooms, satellite dishes, cables, as well as data transfer. |
| July 2006 – Nov 2006 | | Looking for a job vacancy |
| Oct 2002 – July 2006 | | Doing my bachelor's degree in computer science |
| June 2002 – Oct 2002 | | Waiting to start my university study |
| Nov 2000 – June 2002 | | IT technician at 'Bright Star' for computer service & repair. |

| FULL-TIME JOBS | |
|---|---|
| 2018 | • **Project title: Customer Happiness Index (role: data scientist)** <br><br> Supporting Celcom (mobile company in Malaysia) in designing the Happiness Index together with Lynx Analytics, a Singapore - based consultancy, to capture and display business impact of individual product experiences. The goal of Customer Happiness Index was to address this very issue: making sure that the sum of customer interactions and experiences amount to a truly satisfying holistic experience, which in turn translates into lower churn rates and increased revenues from new customers. <br><br> • **Predicting Mobile Number Portability (MNP) among Robi subscribers (role: data scientist)** <br><br> Working with Robi (mobile company in Bangladesh) to examine the effect of MNP on consumers' switching intentions and to build a prediction model for Robi to minimize the impact of MNP. MNP technology enables mobile telephone users to retain their mobile telephone numbers when changing from one mobile network carrier to another. Retaining existing customer has been considered to be one of the most critical challenges for telecommunication service providers than for attracting new one. <br><br> • **Celcom CDR data analysis** <br><br> A Call Data Record (CDR) is an information structure storing applicable information about a given telephonic activity including a customer of a telephonic framework. Each record contains various attributes of the call, such as time, duration, completion status, source number, and destination number. The goal of this project was to process Celcom (mobile company in Malaysia) call records and perform an in-depth analysis, which is helpful for different purposes. |

| PART-TIME JOBS <br> (2015 – 2016) | |
|---|---|
| 2016 | • **Project title: "Big Data for social media analysis" (January 2016 – June 2016)** <br> The aim of the project was to build a Big Data platform that is able to analyze the political opinions of individuals and groups on social media platforms. This would help policy makers in the improvement of legal-making processes. The proposed work included reading data from a social media platform, storing the data in a database, analyzing the data to provide insights, and finally presenting the results as reports and graphs (the project was completed in cooperation with UTM Big Data Center). |
| 2015 | • **Project title: "Muslim Scholar Database" (July 2015 – Dec 2015)** <br> The goal of the project was to build a Muslim scholar database to satisfy researcher needs. The database stored the biographies of Muslim scholars and their relationships. The platform would help researchers from different backgrounds to easily find the information they are looking for and to see how that bit of information (e.g. about a scholar or a specific year) is connected or related to other information in the database. This would eventually help build a network of entities with several dimensions: place, time, and direction of relatedness (the project was completed in cooperation with UTM Big Data Center). |
| 2015 | • **Project title "Social media for disaster recovery" (Dec 2014 – April 2015)** <br> The aim of the project was to enhance communications between individuals during disaster time. The proposed work included the use of social media platforms (e.g. Twitter and Facebook) for collecting information from individuals. It also included the development of a special application that works on top of the current social media platform that is able to collect and transfer data from users to a data collection center for analysis (the project was completed in cooperation with UTM Big Data Center). |

| CERTIFICATES & <br> TRAINING COURSES | |
|---|---|
| 2019 (May) | Certificate of Appreciation: in recognition of services to the academic community in |

| | | |
|---|---|---|
| | | Iraq by serving on the Committee Board of the reviewing to (SAC-SUC2019) 3rd Scientific Annual Conference of Al-Salam University College |
| 2018 (Oct.) | | Applied Social Network Analysis In Python |
| 2018 (Sep.) | | Applied Machine Learning with Python |
| 2018 (Aug.) | | Applied Plotting, Charting and Data Representation in Python |
| 2018 (Sep.) | | Applied Text Mining in Python |
| 2018 (Aug.) | | Introduction to Data Science in Python |
| 2017 | | Strategies on Becoming a Productive Writer (certificate of appreciation) |
| 2017 | | 1st International Conference on Engineering and Computing (ICEC2017) |
| 2016 | | Short Course on Data Science for Beginners (certificate of appreciation) |
| 2016 | | Professional Certificate on Data Science |
| 2016 | | 2nd International Workshop on Big Data Analytics (certificate of appreciation) |
| 2016 | | Three Minute Thesis Competition (3MT) – Faculty Level (certificate of appreciation) |
| 2016 | | Successfully Publishing a Paper Titled "Social Network and Information Dissemination for Disaster Risk Management" (certificate of appreciation) |
| 2016 | | Successfully Publishing a Paper Titled "Predicting the Relevance of Search Results for E-commerce Systems" (certificate of appreciation). |
| 2015 | | International Workshop on Big Data Analytics 2015 (certificate of attendance and certificate of appreciation) |
| 2015 | | A Workshop on Data Science (certificate of appreciation) |
| 2015 | | 2 Days Training on Introduction to Data Science (certificate of completion) |
| 2015 | | Instructor/Speaker at A Web Crawling Workshop (certificate of appreciation) |
| 2015 | | A Workshop on "Self-Organizing Neural Network for Clustering, Classification and Production" (certificate of appreciation) |
| 2015 | | 11th Postgraduate Annual Research Seminar (PARS) |
| 2015 | | Successfully Publishing a Paper Titled "Flight MH370 Community Structure" (certificate of appreciation) |
| 2014 | | 10th Postgraduate Annual Research Seminar (certificate of recognition) |
| 2014 | | Soft Computing Seminar (SOCOS 2014) in Conjunction with Postgraduate Annual Seminar Research (PARS 2014) |
| 2014 | | Internet-based TOEFL Test (total score = 95) |
| 2014 | | 10th Postgraduate Annual Research Seminar (PARS) |
| 2014 | | A workshop on GPU Computing Using CUDA Programming for Machine Learning Algorithms |
| 2012 | | Paper-based TOEFL Test (total score = 603) |
| 2011 | | Advanced Web Technologies |
| 2010 | | CISCO CCNA (institutional course) |
| 2008 | | CCNA Exploration: Routing Protocols and Concepts (academic track) |
| 2008 | | CCNA Exploration: Network Fundamentals (academic track) |
| 2006 | | CCNA(1) – Networking Basics (academic track) |
| 2000 | | 2 Months workshop on How to Use Microsoft Windows 1998 and Microsoft Word 2000 |
| 2000 | | Three-week Workshop on Using Computer Technologies |

| PUBLISHED STUDIES | | |
|---|---|---|
| 2019 | | "Online data preprocessing: a case study approach" International Journal of Electrical and Computer Engineering (IJECE)Vol.9, No.4, August2019, pp. 2620~2626ISSN: 2088-8708, DOI: 10.11591/ijece.v9i4.pp2620-2626 |

| 2018 | "Understanding expert finding systems: domains and techniques", Social Network Analysis and Mining, volume 8, issue 57, publisher: Springer |
|------|------|
| 2017 | "Apache Spark and cluster analysis for expert finding", Int. J. Knowledge Engineering and Data Mining, Vol. 4, Nos. 3/4, 2017. |
| 2017 | "Successful Data Science Projects - Lessons Learned from Kaggle Competition", Kurdistan Journal for Applied Research, Volume 2, Issue 3, August 2017. |
| 2017 | "A Self-Organizing Communication Model for Disaster Risk Management" with Aida Ali. International Journal of Advances in Soft Computing and its Applications (IJASCA), Vol. 9, No. 2, July 2017 |
| 2015 | "Predicting the relevance of search results for e-commerce systems" with Siti Mariyam Shamsuddin and Joel Pinho Lucas. International Journal of Advances in Soft Computing and its Applications (IJASCA), Vol. 7, No. 3, November 2015 ISSN 2074-8523 |
| 2014 | "Flight MH370 Community Structure" in Int. J. Advance. Soft Comput. Appl., Vol. 6, No.2, July 2014. |
| 2014 | "Visualization of Explanations in Recommender Systems" Proceedings of the 2014 2nd International Conference on Innovation and Information Management (ICIIM) – Thailand 2014. |
| 2013 | "Factors Disrupting a successful Implementation of E-Commerce in Iraq" in the Magazine of Baghdad University College of Economic Sciences. |
| 2013 | "E-Government: Latest Trend and Future Perspective the Iraq Case" in the European Journal of Scientific Research, EJSR Vol. 99 Issue 2. |
| 2013 | "Explanations in Recommender Systems Overview and Research Approaches" Proceedings of the International Arab Conference on Information Technology (ACIT) - Sudan 2013. |
| 2012 | "Applying Social Network Analysis to Analyze a Web-Based Community" in 'International Journal of Advanced Computer Science and Applications (IJACSA)' Vol. 3, No.2, 2012 |

| BOOK CHAPTERS | |
|------|------|
| 2017 | "Enhanced Online Service Quality: The Contribution of Local Search Engines". in "Innovations in E-Systems for Business and Commerce", Kadri, El-Hami, eds., Apple Academic Press (ISBN: 978-1-77188-564-5) |
| 2016 | Social Networks and Information Dissemination for Disaster Risk Management (Mohammed Zuhair Al-Taie and Siti Mariyam Shamsuddin). A chapter in E-Systems for the 21st Century: Concept, Developments, and Applications (Kadri, El-Hami, eds.), Apple Academic Press, Volume 1: E-Commerce, E-Decision, E-Government, E-Health, and Social Networks (ISBN 9781771882552) |

| BOOKS | |
|------|------|
| 2017 | "Python for graph and network analysis", Advanced Information and Knowledge Processing, Springer International Publishing AG 2017 (ISBN: 978-3-319-53003-1). |
| 2014 | "Social Network Analysis an Introduction with an Extensive Implementation to a Large-Scale Online Network Using Pajek" by Bentham e-books: http://www.benthamscience.com/ebooks/9781608058181/index.htm |

| CONFERENCES | |
|------|------|
| 2018 | Digital Transformation Asia. Digital Transformation World is TM Forum's flagship event brings together the global Communication Service Provider and telecoms supplier communities in an engaging and innovative event that helps to accelerate the pace of digital transformation (13-15 Nov 2018) |
| 2017 | The 1st International Conference on Engineering and Computing (ICEC2017) held on 22-24 August 2017 in Sulaimani, Kurdistan, Iraq by Sulaimani Polytechnic University (SPU) in partnership with CUS. |
| 2016 | The second international conference on big data analytics, Johor, Malaysia. |
| 2015 | The first international conference on big data analytics, Kuala Lumpur, Malaysia. |
| 2013 | The fourth scientific conference of Baghdad College of Economics Sciences |

| SERVICE TO COMMUNITY | |
|------|------|
| Journal Reviewer | Peer review helps validate research, establish a method by which it can be evaluated and increase networking possibilities within research communities. I have volunteered |

| | | |
|---|---|---|
| | | as a reviewer for several journals in computer science and information technology including "*Social Network Analysis and Mining (SNAM)*" and "*Health Information Science and Systems*", both published by Springer. |
| Book Reviewer | | Reviewing books includes writing a report or essay that offers a critical perspective on a text. Occasionally, I write reviews for new books published by *Manning Publications*. |
| Translation | | Translation includes transforming a text, probably with millions of words, from one language to another. As member of *NASA-In-Arabic* project, we want to keep Arabic readers updated with recent advances in physics, medicine, technology, energy and environment, astronomy, and dentistry. |

**TECHNICAL SKILLS AND COMPETENCES**

| | | |
|---|---|---|
| Programming Languages | • R<br>• Python<br>• Java<br>• MATLAB<br>• C &C++<br>• AJAX<br>• JavaScript<br>• PHP5<br>• SQL, T-SQL | • Extensible Markup Language (XML)<br>• HTML, XHTML and CSS<br>• OpenGL<br>• Visual Basic 6<br>• Prolog<br>• Pascal<br>• MS-DOS<br>• GW-BASIC |
| Applications Familiar with | • SAS Enterprise<br>• Tableau<br>• PyCharm<br>• EndNote<br>• Joomla 1.5<br>• Microsoft Visual Studio 2008.<br>• Packet Tracer.<br>• Eclipse (Java IDEs). | • Maya Desktop.<br>• Corel Draw X4.<br>• AutoCAD<br>• Microsoft Office (Word, Excel, Access, PowerPoint, FrontPage, Outlook Express). |

**LANGUAGES**

| | | |
|---|---|---|
| **MOTHER TONGUE** | | [ Arabic] |
| **OTHER LANGUAGES**<br>• **Reading skills**<br>• **Writing skills**<br>• **Verbal skills** | | [ English]<br>• Excellent<br>• Excellent<br>• Excellent |

**PERSONAL SKILLS AND COMPETENCIES**

- Team leader.
- Install and perform minor repairs to hardware, software, and peripheral equipment, following design or installation specifications.
- Set up equipment for employee use, performing or ensuring proper installation of cable, operating systems, and appropriate software.
- Maintain record of daily data communication transactions, problems and remedial action taken, and installation activities.
- Read technical manuals, confer with users, and conduct computer diagnostics to investigate and resolve problems and to provide technical assistance and support.
- Confer with staff, users, and management to establish requirements for new systems or modifications.
- Develop training materials and procedures, and/or train users in the proper use of hardware and software.
- Refer major hardware or software problems or defective products to vendors or technicians for service.

| Index | Title | Details |
|-------|-------|---------|
| 1 | **Community detection** | **GOAL**:<br><br>The goal is to analyze CDR data using SNA methods<br>  1) Checking the current literature on how to find user communities in social networks.<br>  2) Trying several algorithms for community detection<br>  3) Implementing Louvain algorithm to find communities using the data provided<br>  4) Implementing Louvain algorithm using an application called Gephi and the same data set.<br><br>**SECOND: USING INLUENCE FACTOR METHOD**<br>  1) Finding top 100 users in the network based on number and duration of connections<br>  2) Building the communities of top users and analyze them.<br>  3) Modeling top users' communities such that we have only 10 connections per community.<br>  1. 4) Modeling the removal of overlap between communities using simple heuristics (i.e., random method) |
| 2 | **Propagation model** | - We exploit the strength of social relationships to define how churn influence is propagated in social networks.<br>- Our motivation for this research is based on the assumption that users influence each other over phone conversations.<br>- The power of influence is conditional on the duration and number of conversations.<br>- According to this assumption, a user who has many friends who have already churned is also more likely to churn.<br><br>**METHODOLOGY**<br><br>- Tie-strength is defined based on call duration and call count.<br>- We employ a diffusion model for churn influence to enhance our prediction accuracy<br>- Our work is based on the assumption that once a strong social connection is manifested, there is a high probability that an "idea" (such as churning) will be transferred through the edge from one subscriber to another.<br>- we focus on predicting the probability to churn rather than just labeling the subscriber as a potential churner or non-churner so that subscribers can be ordered from high to low churn likelihood.<br>- A retention campaign can target a limited number of subscribers with the highest likelihood to churn.<br>- We use networkx, pandas and matplotlib libraries |

| | | **Influence propagation model:** |
|---|---|---|
| | | - We use this information to define an influence propagation model where we quantify how influence travels from a churner to his social circle and how much influence is retained by a recipient. |
| | | - There are many models for information flow in a network. |
| | | - We consider a receiver-centric model for influence propagation wherein the receiver of the influence decides how much of the influence to retain. |
| | | - If the receiver is a close friend of the sender, reflected in a higher tie-strength than some other neighbors, then, it is natural to assume that he will be influenced by the senders' actions more than a random friend or colleague he may have social relationships with. |
| | | **Details of our proposed influence propagation model:** |
| | | 1. A churner initiates the spread of influence to his social circle immediately following his decision to churn. |
| | | 1. As soon as a user receives any amount of influence, he replicates that influence and will pass the same amount of influence to all of his neighbors. |
| | | 2. A user who receives influence will decide what proportion of influence to retain. The influence retained is proportional |
| | | 3. to the tie-strength of the link on which the influence is received to the sum of the tie-strengths of all his incoming links. |
| | | 4. A user will not be influenced about the same churn event from the same user twice. |
| | | 5. A user, however, can be influenced by many churn events originating at different nodes. |
| | | 6. Influence will propagate to a maximum of three hops, beyond which the influence to churn is not meaningful. |
| | | 7. 7. The total churn influence accumulated at a node is the sum of influences received at that node, across all churn events. |
| 3 | **Churn model** | - Churn model using SNA |
| | | - We used Interaction Tables: churners with churners, churner with non-churners, non-churners with non-churners. |
| | | - We considered 90 days of no activity as churner. |
| | | - We used SNA to connect users and extract their network structural features. |
| | | - Our work was based on a paper "Social Ties and their Relevance to Churn in Mobile Telecom Networks" |
| | | - We applied data exploration (EDA) |
| | | - We applied data pre-processing |
| | | - The major libraries were networkx and pandas |
| | | - We used "Decision Trees" and "Random Forest Classifier" for prediction. |

| | | **USAGE FEATURES: These include:** |
|---|---|---|
| | | 1) Total Call Frequency |
| | | 2) Number of outgoing calls |
| | | 3) Number of incoming calls |
| | | 4) Total Call volume (seconds) |
| | | 5) Total Incoming call duration (seconds) |
| | | 6) Total Outgoing call duration (seconds) |
| | | 7) Number of unique incoming and outgoing edges |
| | | 8) Number of unique neighbors |
| | | 9) Number of incoming, outgoing calls to/from a different operator's network |
| | | 10) Total Incoming call duration from a different operator's network (seconds) |
| | | 11) Total Outgoing call duration to a different operator's network (seconds) |
| | | 12) Total Incoming, Outgoing edges to/from a different operator's network |
| | | 13) Total number of neighbors in a different operator's network |
| | | 14) Call volume percentage (w.r.t total) to/from a different operator's network |
| | | 15) Call frequency percentage (w.r.t total) to/from a different operator's network |
| | | |
| | | **Interconnectivity Features: These features include** |
| | | 1) Number of adjacent pairs in the set of churner friends i.e., connected by an edge |
| | | 2) Number of pairs in the set of churner friends connected by a path length of 2 |
| | | 3) Number of pairs of churner friends connected by a path that only includes churners |
| | | 4) Total call frequency on edges connecting adjacent churner friends |
| | | 5) Total call volume on edges connecting adjacent churner friends |
| | | |
| | | **Connectivity Features: These features include:** |
| | | 1) Total call frequency to/from churner neighbors |
| | | 2) Total call volume to/from churner neighbors (seconds) |
| | | 3) Call volume and frequency percentages (w.r.t total) to/from a churner |
| | | 4) Maximum call volume, frequency to/from a churner neighbor |
| | | 5) Number of unique churner neighbors |
| | | 6) Percentage of churner neighbors (w.r.t total neighbors) |
| | | 7) Number of non-churner neighbors who have churners as neighbors |
| | | 8) Maximum call volume and frequency with any of these non-churners |
| | | 9) Call volume and frequency percentages (w.r.t total) to/from these non-churners |
| 4 | **Next Best Offer** | - We used Market Basket Analysis |
| | | - We tracked the behavior of customers |
| | | - We divided customers into segments based on their behavior (5 segments) |
| | | - For each segment, we extracted frequent itemsets using the apriori algorithm (from mlxtend library) |
| | | - We extracted association rules from frequent itemsets using association_rules algorithm (from mlxtend library) |
| | | - Because NBO is unsupervised, we used k-means to cluster customers. |
| | | - For each segment, we build next best offer |
| | | - - Important terms: antecedents, consequents, support, confidence, lift, leverage, conviction |

| 5 | **Spark Graphframes** | - Community detecting using Big Data<br>- We used Pyspark and Graphframes<br>- Graphx (a new component in Spark for graphs and graph-parallel computation) was not available under our current Spark distribution<br>- GraphFrames is a package for Apache Spark which provides DataFrame-based Graphs.<br>- I used my local machine<br>- To find communities, I used Label Propagation Algorithm<br>- The algorithm gave us very small and very large community<br>- We used data like this: source, destination, relation (weight)<br>- - The work was successful up to 50m rows |
| --- | --- | --- |
| 6 | **NLP Arabic model** | - The goal is to provide a categorization for customer comments<br>- It is a type of Multi-Label Classification problem<br>- Customer comments are in Arabic<br>- Data is a survey<br>- I used data pre-processing, object standardization, stemming, label encoding<br>- I used Logistic Regression model for prediction<br>- Logistic Regression was the winner against several multi-label classification and deep learning methods<br>- Accuracy started from 64% and reached over 90%<br><br><br>**Data Preparation:**<br><br>1) Dataset cleaning<br>- Remove comments that are not in Arabic and those that do not have a proper categorization.<br>- Data cleaning should be done before any further processing starts.<br>- we collect data from several places. Category distribution of each dataset can be different.<br>- we combine column names of the different datasets.<br>- Merging datasets is done using pd.concat() function.<br><br><br>2) Category consolidation<br>- Category consolidation is about combining the comments that belong to several sub-categories into one main category. The list of categories and sub-categories we received from the source.<br>- For example:<br>- Comments that are talking about: Agent, IVR, Agent, General, Reachability, and Issue resolution ===> Customer Care Issue.<br>- Comments that are talking about: Social Bundles, Balance Loan, Sales, P2P, etc. ===> Service<br>- We update labels because we received a new list of labels from the source (it took them two weeks to update their labels).<br>- We also applied category correction because the labels that we received had different spelling schemas. Therefore, we needed to unify them. |

3) Handling duplicates
   - Duplicate rows are two or more rows that are identical. It is important to remove duplicate rows as they may cause the model to give inaccurate results when overweighting certain terms.
   - We remove duplicates based on two columns. This is to ensure that we keep the comments that have more than one label.
   - We make sure that we remove spaces (strip()) one both sides to remove extra spaces.
   - We also removed duplicates within one cell (duplicates within one cell do not always happen).

4) Combine consecutive rows
   - This will allow to combine the comments that are talking about more than one into one. We combine the labels that belong to the same comment into one column. This is an important step for the next processing steps. We separate the labels using commas (,).

For example:
(prices) <=== الاسعار غالية والاشارة ضعيفة -
(voice-network) <=== الاسعار غالية والاشارة ضعيفة -

After combining consecutive rows:
prices, voice-network <=== الاسعار غالية والاشارة ضعيفة

**Building the prediction model:**

Our current problem is a type of multilabel (or multi-class) classification problems. It a supervised learning problem where an instance may be associated with multiple labels.
For example:
prices, voice-network <=== الاسعار غالية والاشارة ضعيفة
Customer Care issue, Data-network ==> الموظف لا يجيب على الاسئلة والانترنت ضعيف

We experimented with several techniques, including:
1. Machine learning techniques: Logistic Regression and Random Forest, Decision Trees, Naive Bayes
2. Deep learning techniques: Simple Neural Network, Simple Convolutional Neural Network, Multi-Layer Convolutional Neural Network, and Recurrent Neural Network.

| 7 | **Topic Discovery** | **Unsupervised learning with social media:** |
|---|---|---|
| | | |
| | | - The goal is to find topics in user comments on FB (what are they talking about) |
| | | - For this purpose, we used unsupervised methods like topic modeling (e.g. LDA algorithm) |
| | | - We also used sentence clustering (with the help of k-means) |
| | | - We also used sentence similarity (like Fuzzy Similarity: FuzzyWuzzy is a library of Python which is used for string matching) |
| | | - We also used WordCloud. |
| | | - We didn't achieve good results with this |
| | | |
| | | **Semi-Supervised learning with social media:** |
| | | |
| | | - For this purpose, we used Anchored-CorEx algorithm |
| | | - Anchored CorEx is an extension of CorEx that allows the "anchoring" of words to topics. |
| | | - We provide a set of anchor words for each category |
| | | - Anchor words should be unique and representative |
| | | - We fit the CorEx model (with anchors) to our dataset |
| | | - We achieved success of 70% |
| | | - - We also used Guided-LDA algorithm but gave no good results. |
| 8 | **MBB Propensity model** | - The purpose of this model is to provide a more advanced tool to better identify the best customers to target in order to drive MBB sales. |
| | | - This propensity model aims to assign customers a % score that defines their likelihood (or "propensity") to buy a MBB device. |
| | | - We defined our data based on a survey call, where customers either reply to the call or not, and whether they agree to buy the product or not |
| | | - The data was un-balanced because the rate of successful to the overall calls is very low |
| | | - This problem is a type of rare event problem |
| | | - We used several features for the model, related to customer previous expenditure behavior |
| | | - We calculated feature importance to filter out un-necessary features |
| | | - Top features were voice revenue, VAS revenue, data revenue, top-up count. |
| | | - We used several classifiers like random forest and logistic regression. LR was better |