IMPROVING PHARMACOVIGILANCE SYSTEMS USING NATURAL LANGUAGE PROCESSING ON ELECTRONIC MEDICAL RECORDS
## PROJECT TEAM

<table>
<thead>
<tr>
<th>Names</th>
<th>Institution/Co</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms. Gloriana J. Monko</td>
<td>UDOM</td>
</tr>
<tr>
<td>Mr. Steven Edward</td>
<td>UDOM</td>
</tr>
<tr>
<td>Mr. Ibrahim Mtandu</td>
<td>Capital Space</td>
</tr>
<tr>
<td>Mr. Zephania Reuben</td>
<td>UDOM</td>
</tr>
<tr>
<td>Mr. Waziri Shebogholo</td>
<td>UDOM</td>
</tr>
</tbody>
</table>
• This research focuses on enhancing Pharmacovigilance Systems using Natural Language Processing (NLP) on Electronic Medical Records (EMR).

• The major task is to develop an NLP model for extracting Adverse Drug Reaction (ADR) cases from EMR.
Problem tree of the study

**Long term Effects**
- Increase of Morbidity & Mortality
  - Patient lose confidence towards physicians
  - Uninformed decision on the monitoring and control of drugs
  - Prolongs stay in healthcare facility
  - Necessitates admission to hospital

**Immediate Effects**

**Central Problem**
- Under Reporting of ADR
  - Lack of Feedback After Reporting
  - Lack of sufficient Knowledge to Report
  - Lack of Staff Motivation

**Key Contraints**
- Insecurity
- Complacency
- Diffidence
- Indifference
- Ignorance
- Lack of Time to Report ADR

**Root Causes**
- Fear of Medical Consequences
Proposed Solution

ADR FRAMEWORK
The dataset for the research was gathered from health care databases, where the identified areas were two hospitals and data from a public health database i.e.

BM Hospital

MIMIC III

UDOM Hospital
• We were able to obtain records of the year 2017, 2018, and 2019. The complete dataset has 22210 rows.

• There were challenges faced while working on these data since they were not clear or systematically reported on tracking the complete patient history which is one of the important features in tracing all the events occurring during medication that can help in identification of ADR case(s).

• Combined columns; History, Allergy, Diagnosis and Medication to create **text documents** which can give a certain logical patient's history per each data point.
BM Hospital Dataset

- From BM hospital we were able to extract 600 pdf reports from the hospital system, in which a single report comprised of information such as:
  - Progress Notes
  - Diagnosis
  - Medication Order
  - Lab Notes

- Annotation
- Pre-processing
MIMIC-III Dataset

(Source: Johnson et al., 2016)
MIMIC-III Dataset

• Due to the challenges faced while analyzing the databases from two mentioned hospitals, we firstly focused on MIMIC-III clinical care database to create a baseline model.

• In MIMIC-III database we were able to extract a table named noteevents which contains the complete patient history from admission to discharge.

• We queried 100,000 records from the table, in the text column where there were information on prior patient history, medication, dosage, examination change in medication etc.
Methods & Results

• We annotated and performed data analysis and cleaning techniques in MIMIC-III data such as sampling and regular expression.

• We filtered the text documents based on keywords like adverse, reactions, reaction, adverse events and adverse drug reaction.

• 3446 rows of records contain words which guided the team in labeling process. Finally Non ADR cases and ADR cases labeled to 0 and 1 respectively.

• NLP approaches have been used to create a baseline model using MIMIC-III clinical care database, where we applied techniques such as tokenization, lemmatization, stemming and stop word removing. Also we applied TF-IDF in encoding the text documents.
Methods & Results

• In modeling we worked with classical machine learning classification algorithms using scikit-learn.
• We trained six different models which are Support Vector Machine, eXtreme Gradient Boosting, Gradient Boosting, Adaptive Gradient Boosting, Multilayer Perceptron and Random Forest

• The models accuracy scores are summarized below:-

<table>
<thead>
<tr>
<th>Model name</th>
<th>Accuracy Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machines</td>
<td>0.87546</td>
</tr>
<tr>
<td>eXtreme Gradient Descent</td>
<td>0.853611</td>
</tr>
<tr>
<td>Adaptive Gradient Boosting</td>
<td>0.840617</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.870594</td>
</tr>
<tr>
<td>MultiLayer Perceptron</td>
<td>0.885424</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.878717</td>
</tr>
</tbody>
</table>
Methods & Results

• In model validation we optimized some of the models to improve the performance, the optimized models are Support Vector Machine, MultiLayered Perceptron and Random Forest.
• The results after optimization are summarized below:-

<table>
<thead>
<tr>
<th>Model name</th>
<th>Accuracy Score</th>
<th>Precision</th>
<th>recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>MultiLayered Perceptron</td>
<td>0.90</td>
<td>0.90</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Classification Report and Confusion Matrix for Support Vector Machines Classifier
Methods & Results
Classification Report and Confusion Matrix for MultLayer Perceptron Classifier
Methods & Results
Classification Report and Confusion Matrix for Random Forest Classifier

CM for RFClassifier Tuned

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>225</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>196</td>
</tr>
</tbody>
</table>
Conclusion & Future Work

• Based on the results obtained from the baseline model the best model choice was Random Forest because it was having lower numbers of false positive and false negative compared to other optimized models.

• One of our future work is to try working with word embeddings and other deep learning techniques and compare the results to those of classical machine learning techniques.

• The BM dataset preparation is underway and we looking forward to finish as soon as the pharmacologist is done with annotation.