**mWACH Analysis of Patient Engagement using Natural Language Processing techniques**

# Introduction

This data science internship project investigated the value of Natural Language Processing (NLP) techniques for a prediction task in Maternal/HIV mHealth. The Mobile WACh NEO text message database contains data on more than 25,000 text messages sent to pregnant, HIV+ women in Kenya in the treatment arm of an RCT of 2-way SMS messaging to encourage adherence to antiretroviral therapy (ART).

The Mobile WACh study, led by Jenn Unger and Keshet Ronnen, have been investigating the association between engagement and antiretroviral adherence prediction. As a complimentary analysis to this work, this project set out to measure the ability of machine learning methods to predict engagement. i.e. how likely is it that a text message from the system will yield a response?

# Problem Statement

There is need to know whether Machine Learning approaches in Natural Language Processing can help in predicting engagement and which models perform better than others?

# Project Objectives

1. **General Objective**

To measure the ability of NLP machine learning methods to predict engagement.

1. **Specific Objectives**
   1. To measure the prediction power of various NLP methods and understand their contribution to predicting engagement
   2. To evaluate whether NLP methods can be used to compliment prediction methods like logistic regression and gradient boosting machines to increase rate of prediction.

# Methodology

Previous work has been done by the team (Jenn and Keshet), using factors like time of message, age of participant, and language preference to predict engagement, using biostatistical methods like logistic regression, and have experimented with non-parametric alternatives like Gradient Boosting Machines (GBMs).

This follow up work extended the logistic regression and GBMs to use word/document embeddings of the text message itself. The aim was to quantify the degree to which the content of the text message might impact engagement.

The following models were applied in this analysis: -

1. Bag-of-words and n-gram models, where the English-translated text message was encoded based on words and used as additional predictors in machine learning prediction methods, like GBM or penalized logistic regression.
2. Word2Vec/Doc2Vec models, where instead of a simple n-gram encoding, the English-translated content of the text is embedded in a low-dimensional vector space using a pre-trained neural network approach, such as Doc2Vec.
3. Fine tuning a BERT model. BERT, stands for Bidirectional Encoder Representations from Transformer, is a bidirectional model that considers the position of the words with respect to their location in a paragraph. It makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.

**Data**

The project used SMS data from the mWACH-NEO database that had over 59,000 messages, of which over 25,000 were sent by the system. Apart from the text messages themselves, the following variables were also used as part of feature vectors: -

* edd\_wk – time to expected date of delivery in weeks.
* study\_wk – Duration of stay since enrolment date in weeks
* past\_response – A Boolean value for past response of previous message
* native\_language – A Boolean value for English (0) or non-English (1) langauge that the message was sent on.
* delivery\_status – Boolean outcome based on system status of whether message was delivered to participant (1) or not (0).
* swahili, english, luo – Language the message was sent in.

**Training & Evaluation**

The analysis tools used were based on Python programming language and the environment used was a Jupyter notebook.

Data was prepared for analysis by extracting the features into a feature matrix X and a response vector Y. The data was split into training and test sets in the ratio 80:20 and then the models were fit using the training set and finally tested on the test data.

Model performance was done using two tools namely: -

1. Accuracy Score (accuracy\_score) – A sci-kit learn implementation of accuracy score, which computes the percentage of the set of labels predicted for a sample that exactly match the actual corresponding set of response labels (scikit-learn.org).

2. Area Under the Receiver Operator Curve (roc\_auc) – A sci-kit learn implementation which is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. (Bewick, Cheek, & Ball, 2004).

K-Fold Cross Validation – An extended validation method was used where the train and test datasets were automatically shuffled multiple times and results for both accuracy score and area under the receiver operator curve were taken and their mean computed.

# Results and Discussion

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Model | Features | Accuracy Score | Receiver Operator Curve (AUC) | K-Fold accuracy Score | K-Fold Receiver Operator Curve (AUC) |
| 1 | Logistic regression | edd\_wk,study\_wk,past\_response,native\_langauge,delivery\_status,swahili,english,luo | 68.5828466 | 73.1132322 | 69.38048099 | 73.7819412 |
| 2 | GB Classifier | edd\_wk,study\_wk,past\_response,native\_langauge,delivery\_status,swahili,english,luo | 70.3197049 | 75.1209787 | 71.31030939 | 75.6430781 |
| 3 | Logistic Regression on BoW | BoW | 66.5770265 | 70.7354905 | 66.93057731 | 71.0639032 |
| 4 | Logistic Regression on BoW | BoW + edd\_wk,study\_wk,past\_response,native\_langauge,delivery\_status,swahili,english,luo | 71.8401844 | 78.4645289 | 72.36847867 | 78.6990292 |
| 5 | GB Classifier on BoW | BoW | 64.3804035 | 67.5927898 | 64.09631031 | 66.7512647 |
| 6 | GB Classifier on BoW | BoW + edd\_wk,study\_wk,past\_response,native\_langauge,delivery\_status,swahili,english,luo | 70.3197049 | 75.1209787 | 71.71515606 | 77.4714086 |
| 7 | Logistic Regression on Bi-grams | Bi-grams only | 67.2622 | 72.3921 | 67.6435 | 72.0577 |
| 7 | Logistic Regression on Bi-grams | Bi-grams + edd\_wk,study\_wk,past\_response,native\_langauge,delivery\_status,swahili,english,luo | 72.3150817 | 78.5453406 | 72.4876098 | 78.8361716 |
| 8 | GB Classifier on Bi-grams | Bigrams + edd\_wk,study\_wk,past\_response,native\_langauge,delivery\_status,swahili,english,luo | 71.0470701 | 76.900838 | 71.6959418 | 77.6427391 |
| 9 | Gensim doc2vec on GB classifier | Doc2Vec embeddings | 72.4101 | 76.6189 | 71.382705 | 75.8079 |
| 10 | Gensim word2vec on GB classifier |  |  |  |  |  |

**Discussion**

1. **Logistic regression using a list of features vs using BoW only**

The message texts were encoded using a large corpus and the resulting vectors were used as a set of feature vectors on a logistic regression task and their performance compared to a set of selected input vectors based on characteristics of the message and recipients.

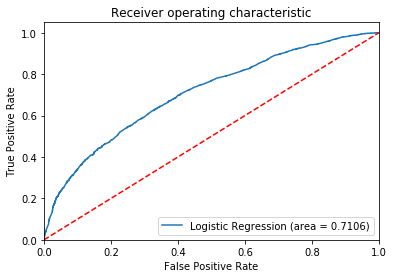
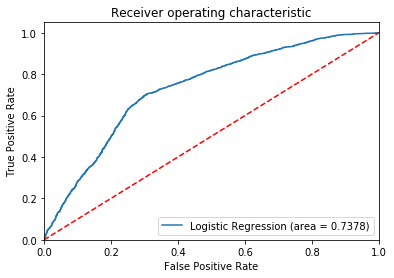


Fig 1. Logistic regression on features only. Fig 2. Logistic regression on BoW only

What was observed is that the BoW method comes so close (a margin of 2% compared to normal feature vectors) to predicting a response, despite its lack of word-order and meaning representation. The overall prediction using either the feature vectors or BoW independently were both recording at around 70% likelihood.

The test was repeated, this time using a bi-gram (two word-length embedding) and the same set of feature vectors. There was a 1% improvement from BoW method, which significantly reduced the margin of error.

1. **Prediction power – Increased prediction power when using NLP models as part of feature vectors**

There was a huge prediction boost that was observed when any of the NLP methods (BoW, Bi-grams, Word2Vec or Doc2Vec) was used alongside a selected set of feature vectors. The below ROC AUC curves demonstrate this prediction power increment.

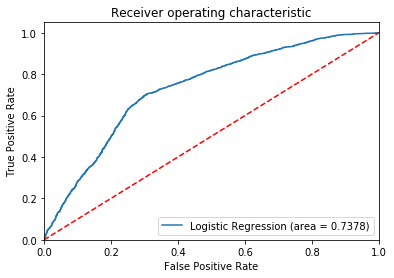
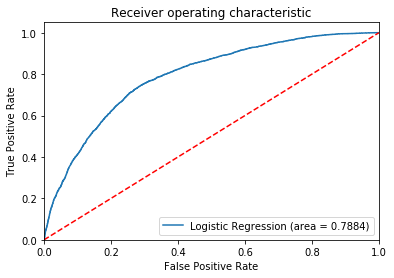
 

Fig 3. Logistic regression on features only. Fig 4. Logistic regression using Bi-grams and selected features

From the above figures, we see an increment of more than 5% when using an NLP method.

Gradient Boosting Machines were also noted to increase prediction power compared to normal

logistic regression models as shown below.

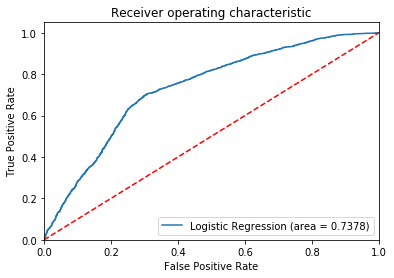
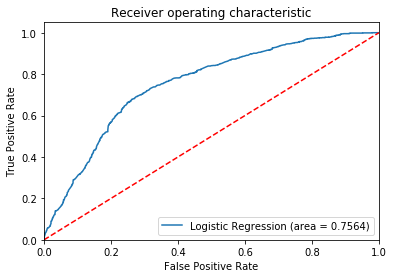
 

Fig 5. Logistic regression on select features Fig 6. GB Classifier on same features

The charts above show an increase in 2% on prediction power when using Gradient Boosting Classifiers compared to normal logistic regression. However, this is not always the case as sometimes using GB classifiers results in a lower prediction rate as seen in the table of results above.

1. NLP for predicting engagements

Finally, we looked at a more advanced NLP model, the Doc2Vec. This is a bidirectional text embedding method that takes into account both position of word and the context of the words in paragraphs. Using Doc2Vec only, without any additional feature vectors delivered a 75% prediction roc\_auc score and 71% accuracy score. The scores were much better than most basic NLP model combinations with feature vectors.

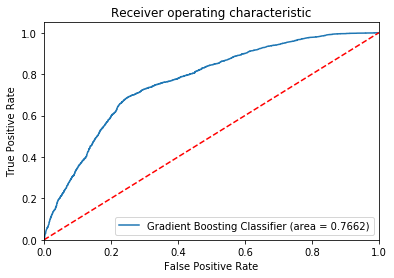


Fig 7. GB Classifier on Doc2Vec feature matrix

Previous analysis cited in literature (Jey Han Lau and Timothy Baldwin, 2016) suggests results should be in the 80% range, however, this exercise could not advance the performance further. Probably more tests will be done in the future.

**Concerns noted**

1. Gensim Doc2Vec performance

Jey and Timothy, in their review suggest that the Gensim Doc2Vec implementation performs better with pre-trained word embeddings. The default setting initializes the word embeddings randomly which contributes to degrading performance in the model according to their findings.

2. Gradient Boosting performing lower than logistic regression on some models

Probably this scenario was caused by overfitting on the logistic regression part. More tests will be conducted in the near future to remedy this outcome.

**Next Steps**

1. Training and evaluating BERT

This project will perform training and testing a BERT implementation based on Google’s recently open sourced BERT model.

2. Revisiting Doc2Vec

The Doc2Vec model will continue to be fine tuned and trained.

# Conclusion

From the results discussed here, it is clear that NLP models can be used to predict occurrence of a response. More advanced methods like Doc2Vec have higher prediction power even on their own without the need to supply extra features.

We have also seen that given a set of pre-selected features, NLP models can compliment these features and increase prediction power.