|  |  |
| --- | --- |
|  | **Data Science Assignment** |
|  | Submitted To  Indegene Private Limited  Submitted by  Abin Singh R  Date  29-10-2018 |

|  |
| --- |
| **[Machine Learning model To classify Adverse drug reaction and Non Adverse Drug Reaction from the twitter text]** |
|  |

**Table of Contents**

**01. Problem Statement**

**02. Data Sourcing, Data Cleaning and Sampling**

**03. Bag-of-Words (BOW) Model**

**04. Cost Sensitive Learning and Evaluation Metrics**

**05. Baseline and Traditional Model**

**06. Implementing Neural Network**

**07. Word Embeddings and Embedding Layer**

**08. Implementing Convolution Neural Network**

**09. Finalizing the Model and Cross-Validation**

**10. Conclusion**

**01. Problem Statement**

Pharmacovigilance identifies ADRs (adverse drug reactions) after a drug’s release. Comprehensive knowledge of ADRs can reduce the detrimental impact on patient’s health system.

Practically, clinical trials cannot investigate all settings in which a drug will be used, making it impossible to fully characterize the drug’s adverse effect profile before its approval.

Pharmacovigilance methods continuously analyse frequently updated data sources, Twitter in particularly because of its large user base, demographic variability, and publicly available data. ADR detection in social media requires automated methods to process the high data volume.

**Analytical Framework**

In-order to solve the above problem, we can develop the supervised machine learning model which can extract the required features from the text and classify the ADR / Non-ADR with better accuracy.

**02. Data Sourcing, Data Cleaning and Sampling**

Pandas and nltk library were used to load and clean the raw data.

a. Tokenization (Word and Regex)

b. Convert Tokens to Lower Case

b. Filter out Punctuation

c. Filter out Stop Words (and Pipeline)

d. Stem Words

**Sample to train and test set**

Sampling is done with 80:20 ratios to keep test set for validating the performance.

**03. Bag-of-Words (BOW) Model**

[CountVectorizer](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html) from Scikit-Learn is used to vectorize the sentences.

**04. Cost Sensitive Learning and Evaluation Metrics**

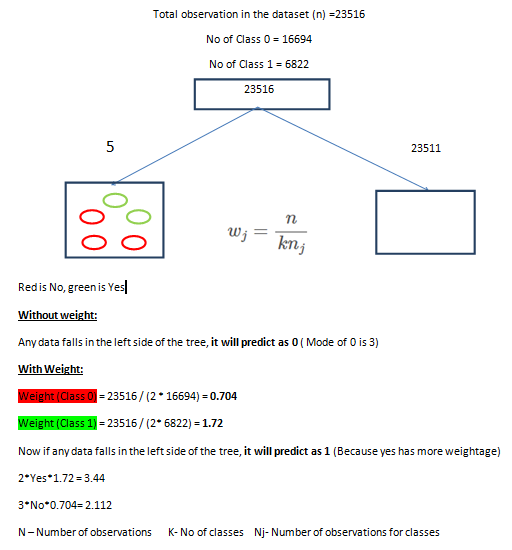


Figure 1: Hand Calculated (For understanding purpose only)

As our dependent variable Adverse Drug Reaction (ADR) Label is biased to the user who tweets about Non-ADR. So, dataset should be processed with undersampling, Oversampling or Cost Sensitive Learning methods.

As Cost Sensitive Learning (CSL) method is an effective method to tackle imbalanced datasets. So, CSL is applied for the traditional ML models.

**Evaluation Metrics**

Our business wants to correctly identify the Adverse Drug Reaction (ADR) and Non-Adverse Reaction (Non-ADR) from the tweets.

Considering false negative rate (Classifying ADR as Non-ADR whose cost is more) and false positive rate (Classifying Non-ADR as ADR), f score (weighted recall & precision) metric is evaluated for the entire model building process.

**05. Baseline and Traditional Model**

Logistic Regression is used as the base-line model to compare with the more advanced models.

Traditional ML models which are mentioned below are also used to test the performance.

1. Random Forest
2. Support Vector Machine
3. Adaptive Boosting
4. Gradient Boosting

**Grid Search Method**

To choose the optimal parameter, Grid Search method is used for the algorithm which has hyper parameter.

Parameter grid values are taken by considering the bias and variance trade-off.

**06. Implementing Neural Network**

To further increase the f\_score, Neural Network model is implemented.

Model is developed with single hidden layer.

With suggestion from various research papers, **Adam** optimizer is choosed for the all the network models

**Library used**: Keras neural network

**07. Word Embeddings and Embedding Layer**

In-order to achieve better f\_score, word embeddings method is implemented in the neural network model.

Word embeddings - To represent each word as dense word vectors

To tackle the variation of each sentence, keras **pad\_sequence** is used.

Keras emedding layer is constructed to map the calculated integer to the embedding space.

**08. Implementing Convolution Neural Network**

Dense vectors from word embeddings are implemented with CNN architecture.

Max pooling and hidden layers with different kernel size are tuned to achieve better accuracy in the validation data.

**09. Finalizing the Model and Cross-Validation**

Performance of CNN Model Performance of SVM

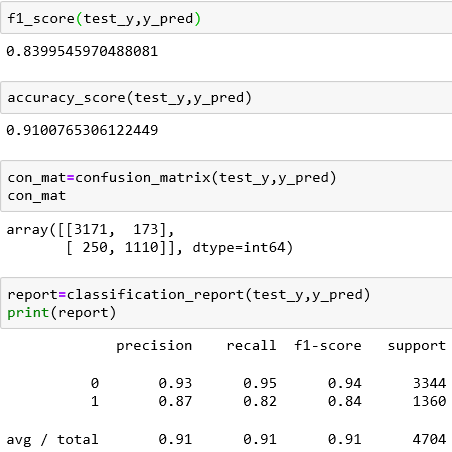
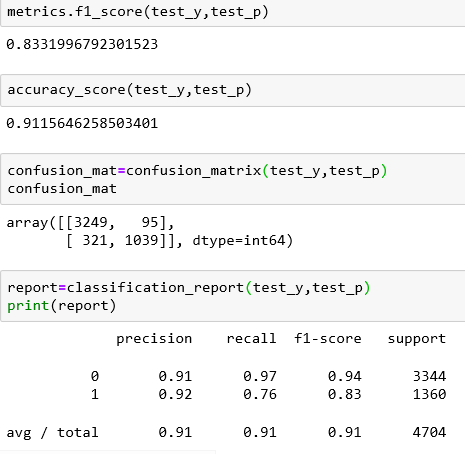
 

Figure2: Extracted from Jupyter Notebook

Actual business problem is to automatically segment ADR and Non- ADR tweets. If we mis-classify ADR as Non-ADR, then the impact towards human health is a question mark. And on the other hand if we mis-classify Non-ADR as ADR, then the clinical procedure to test again the drug in lab and certification cost is also a question mark.

So, in-order to balance the above mentioned problem, model with better f\_score is choosed as the final model.

When compared with all other models, harmonic average of recall and precision (F\_Score) is high for Convolutional Neural Network (CNN), So CNN is choosed as the final model

**Cross-Validation**

Generalization of the model for CNN and SVM is tested and validated.

Considering future if we extract more text from twitter source, CNN model can do the same or more performance as it can handle large datasets when compared with the SVM model.

**Confusion Matrix for final model**

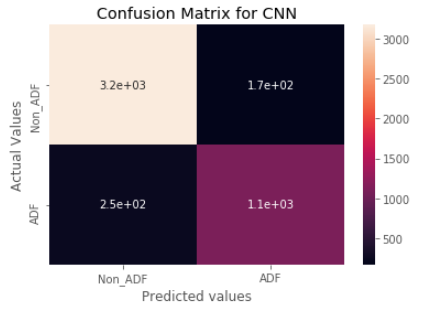


Figure 3: Extracted from Jupyter Notebook

**10. Conclusion**

With this study, we analysed, cleaned the raw twitter data using (Tokenization etc.) and extracted the features with the help of Bags of Model (BOW) and Word Embeddings. And we started with base line and traditional ML models. In an aim to achieve better f\_score metric, we implemented Neural Network and Convolution Neural Network models.

Finally as per our business requirement, classification model for Adverse Drug Reaction and Non-Adverse Reaction from twitter text data with a descent f\_score of 84 (CNN Model) is ready to deploy.