FINAL PROJECT REPORT

ADTrap- Adverse Drug Reaction Classification using NLP tools & Deep Learning Methods

Ritika Saxena- ritikasaxena0320@gmail.com
Kshitij Soni- kshitijsoni@iipe.ac.in , kshitijsoni318@gmail.com

1. Project Goals

Our primary intent is to explore the extent to which ADR assertive text segments can be classified from text based data sources, particularly social media sources and structured datasets. The following list summarizes our intents:

(i) Investigation of different Neural Network (NN) architectures for ADR classification.

(ii) Exploration of NLP techniques to extract informative and portable features from text coming from distinct sources.

(iii) Investigate the performance of supervised classification approaches on data from social media to data from other more structured and unstructured sources.

(iv) Optimize machine learning and deep learning algorithms to improve performance over existing approaches.

2. Related Work

The closest to our work on ADR classification comes from [1] who suggested two new NN models, Convolutional Recurrent Neural Network (RCNN) by concatenating convolutional neural networks with recurrent neural networks, and Convolutional Neural Network with Attention (CNNA) by adding attention weights into convolutional neural networks.

For data-mining we referred to [2, 3] as it is well known that training phase is much difficult and the phase wraps around. Despite the vast amount of information available on social networks, research on mining that data for ADR classification is still very much in its infancy.

We also referred to [4, 5, 6, 7] for training of the data through various methodologies, comparison between models on different types of data and its combinations, and also to understand the various difficulties in it.

3. Datasets

We have used Twitter dataset containing informal language. The tweets associated with the data were collected using generic and brand names of the drugs, and also their possible phonetic misspellings. The tweets were annotated for presence of ADRs. Also sample data used for this predictive modelling was obtained from the United States of America federal drug agency database(ADE Dataset). The database contains reports of drug adverse events which occurred in various countries of the world. A total of 92130 drug adverse reports were extracted from the database with selected 52 attributes of each reported events.

4. Different Approaches

4.1 Data cleaning

The data cleaning process is implemented to drop redundant and duplicated variables that are not required, also create uniform datatypes in each column of dataframe. This process of cleaning each column will remove meta characters, numerical value in text columns and texts from numeric columns. This will produce same data type for all values of a variable. It will increase accuracy of plots, data engineering and modelling. The unique values in each columns will be examined for cleaning columns where necessary, functions are created to clean numeric and objects data types respectively. The count of unique values will be displayed before and after cleaning to check any deviation

4.2 Exploratory Data Analysis and Visualization

The plots are produced using Matplotlib and Seaborn libraries give a differentiation or separation of the samples in the binary classifier. The visualization will show relationship between input features or variables. Visualize data distribution of each variable for skew correction. This will also help to discover trend and patterns in the data and to understand data characteristics. The Analysis is also aimed at discovering relationships in data engineering choice. Plot include univariate plots using Histogram, Barplot, Bivariate plots such as Boxplots, Multivariate scatter plots and cluster plots.

4.3 Model Building and Hyper-parameter Tuning

4.3.1 Feature Selection

Feature Seletion -Recursive Feature Elimination (RFE) repeatedly constructs a model and choose either the best or worst performing features. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

4.3.2 Training and Model Fitting

In the binary classification task, the following models are fitted and compared using different evaluation metrics. Logistic regression, NaiveBias, SVM, RandomForest, XGboost, Gradient Descent parameter optimization.

In both the neural network architectures- Convolutional Neural Network (CNN) and Convolutional Neural Network with Attention (CNNA), the training algorithm is Adadelta (Zeiler,2012) with learning rate of 1.0, decay rate (ρ) of 0.95 using library Keras. The embedding is trained together with other parameters. For each fold, we split the training dataset into training and validating sets. The training stops when there is no performance improvement on the validation set after 5 consecutive epochs. The batch size is set as 50. All convolutional window has a size of 5.

5. Experiments, Results and Conclusion

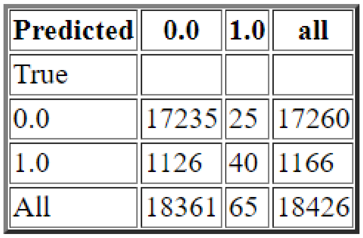
5.1 Code and Environment

We used python language for conducting all our experiments.

5.2 Results

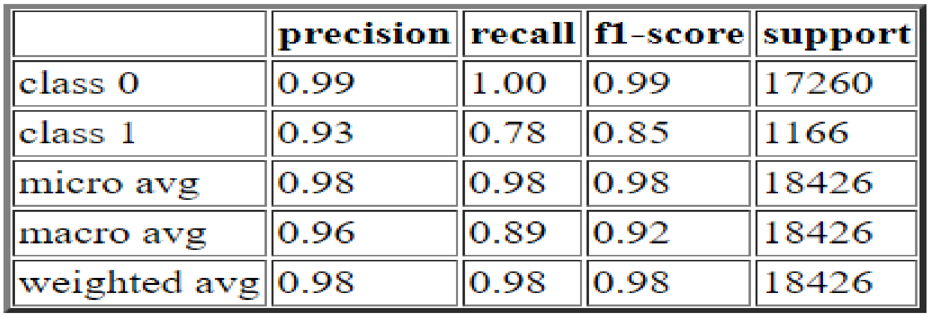
The more data intensive estimators gave better performance precision and recall than logistic regression. Comparison of prediction accuracy by the models shows that logistic regression, RandomForest and K Nearest Neighbor gave the similar performance accuracy based on the data. Accuracy on obtained are as follows:-
Logistic Regression = 0.9361228698578096
RandomForest = 0.9810593726256377
K Nearest Neighbor= 0.968305655052643.

The results from these models shows that with more data, feature engineering and hyper-parameter tuning on RandomForest and KNN, the performance will be improved.



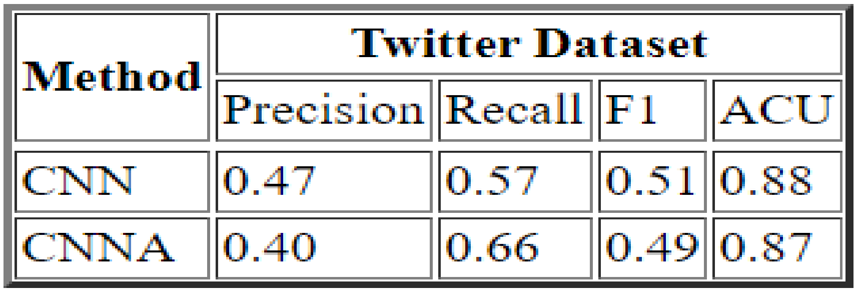
Predicted	0.0	1.0	all
True			
0.0	17235	25	17260
1.0	1126	40	1166
All	18361	65	18426

Table 1. Confusion Matrix on ADE Dataset- Logistic Regression Model



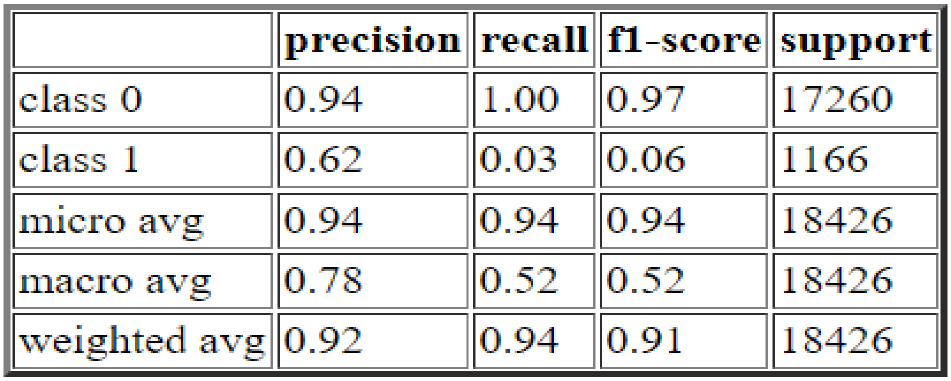
	precision	recall	f1-score	support
class 0	0.99	1.00	0.99	17260
class 1	0.93	0.78	0.85	1166
micro avg	0.98	0.98	0.98	18426
macro avg	0.96	0.89	0.92	18426
weighted avg	0.98	0.98	0.98	18426

Table 2. Adverse Drug Reaction Classification- Random Forest Model on ADE Dataset



Method	Twitter Dataset			
	Precision	Recall	F1	ACU
CNN	0.47	0.57	0.51	0.88
CNNA	0.40	0.66	0.49	0.87

Table 3. Adverse Drug Reaction Classification on the Twitter Dataset



	precision	recall	f1-score	support
class 0	0.94	1.00	0.97	17260
class 1	0.62	0.03	0.06	1166
micro avg	0.94	0.94	0.94	18426
macro avg	0.78	0.52	0.52	18426
weighted avg	0.92	0.94	0.91	18426

Table 4. Adverse Drug Reaction Classification – Logistic Regression Model on the ADE Dataset

We compared the precision, recall and F-scores of the neural network architectures on the Twitter Dataset. We also reported the Area Under the ROC Curve (AUC) results. It can be observed that in general, results on the ADE dataset (logistic regression, Random Forest & KNN models) are better than those on the Twitter Dataset (CNN & CNNA models). This is perhaps not surprising since tweets contain a lot of ill-grammatical sentences and short forms. Simply relying on an ADR lexicon for the detection of ADRs from text gives the worst results. Overall, CNN gives the best results although CNNA are quite close to CNN in terms of AUC values. Our hypothesis is that as

ADR descriptions are composed of short fragments of texts, convolutions with small windows are enough to capture necessary information for ADR classification.

5.3 Conclusion

This project has explored two different neural network (NN) architectures and models like logistic regression, Randomforest, and KNN(k- Nearest Neighbor) for ADR classification. Among NN architectures, no significant differences were observed on the Twitter Dataset. RandomForest model appears to perform best among all the models on the ADE Dataset. Nevertheless, CNNA allows the visualization of attention weights of words when making classification decisions and hence is more appropriate for the extraction of word subsequences describing ADRs.

6. Efforts

6.1 Most Challenging Part

The most challenging part was training the phase of tweets and reports, in accordance with observations in similar previous experiments. Most of our efforts were directed towards reducing this error.

6.2 Fraction of Work Done by Different Members

There was no specific work division. Every member did whatever work was there at hand. All the members equally contributed to the project.

7. References

[1] Trung Huynh, Yulan He, Alistair Willis, Stefan Ruger: Adverse Drug Reaction Classification with Deep Neural Networks. *COLING 2016*

[2] {Harpez et al.2012} R Harpaz, W DuMouchel, N H Shah, D Madigan P Ryan and C Friedman. 2012. Novel Data Mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics, 91(6):1010-1021*

[3] {Ginn et al.2014} Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apur Patki, Karen Oconnor, AbeedSarker, Karen Smith, and Graciela Gonzalez. 2014. Mining Twitter for Adverse Drug Reaction Mentions:A Corpus and Classification Benchmark. *Inproceedings of the 4th Workshop on Building and EvaluatingResources for Health and Biomedical Text Processing (BioTxtM).*

[4] {Sarker and Gonzalez2015} Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53:196–207*

[5] {Yates and Goharian2013} Andrew Yates and Nazli Goharian. 2013. ADRTrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites. *In European Conference on Information Retrieval (ECIR), pages 816–819.*
[6] {Zhang et al.2016} Zhifei Zhang, Jian-yun Nie, and Xuyao Zhang. 2016. An Ensemble Method for Binary Classification of Adverse Drug Reactions From Social Media. *In Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*
[7] {Glorot et al.2011} Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Net-works. *In 14th International Conference on Artificial Intelligence and Statistics (AISTATS), pages 315–323*.