

DRUG ADVERSE EVENTS CLASSIFICATION USING SOCIAL MEDIA CONTENT

Ranith Sachintha Ranawaka

(179346C)

Degree of Master Of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2021

Declaration

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic, or another medium. I retain the right to use this content in whole or part in future works.

.....

Ranith Sachintha Ranawaka

.....

Date

The above candidate has carried out research for the Master's thesis/ Dissertation under my supervision.

.....

Dr. Surangika Ranathunga

.....

Date

Acknowledgements

I would like to express profound gratitude to my advisor, Dr. Surangika Ranathunga, for her invaluable support by providing relevant knowledge, materials, advice, supervision, and useful suggestions throughout this research work. Her expertise and continuous guidance enabled me to complete my work successfully. Further, I would like to thank Mr. Sanmugan Aravinthan, for providing valuable resources for this project.

I am grateful for the support and advice given by Dr. Uthayasanker Thayasivam. Further, I would like to thank all my colleagues for their help in finding relevant research material, sharing knowledge and experience, and for their encouragement.

I am as ever, especially indebted to my parents and family for their love and support throughout my life. Finally, I wish to express my gratitude to all my colleagues at IQVIA Sri Lanka, for the support given to me to manage my MSc research work.

Abstract

On-time detection of possible adverse events a drug may have has been a critical issue for the pharmaceutical industry, although it undergoes rigorous clinical trials there still can be adverse effects once it reaches the market, this is known as post-market drug safety surveillance. The ordinary way to collect these was through physicians who prescribe the drug reporting back to the pharmaceutical company. But this process consumes time and has the risk of missing important drug adverse reactions.

The recent popularity of social media has led people to communicate extensively about their aspects in day-to-day life, this includes the communications of the experience regarding the drugs and their adverse events. This makes social media a rich resource for monitoring drugs after they reach the market.

In this research, we experiment with machine learning models including deep learning models using social media contents manually verified by health care professionals for the presence of drug adverse events. The Social media data has been acquired through popular health care social media channels from their respective APIs.

Well-known Text classification algorithms such as SVM and Logistic Regression provide the best accuracy for ADR mining, CNN's which has recently shown high accuracy levels for text classification also shows high levels of accuracy for ADR classification tasks.

Table of Contents

Declaration	i
Acknowledgements	ii
Abstract	iii
Table of content	iv
List of Figures	vii
List of Tables	viii
List of Equations	x
List of Abbreviations	xi
1.Introduction	1
1.1 Drugs and Clinical Trials	1
1.2 Drug Adverse Reactions	2
1.3 Pharmacovigilance	3
1.4 Problem/Opportunity	4
1.5 Objectives	5
1.6 Contributions	5
1.7 Thesis Structure	6
2.Literature Review	7
2.1 Definition	7
2.2 Data Preprocessing Techniques	8
2.3 Binary classification of ADR and Non-ADR text	9
2.4 Drug and ADR Relationship Extraction	10
2.5 Feature Engineering	12
2.5.1 Word Embeddings	13
2.5.2 BOW (Bag of Words)	14
2.5.3 TF – IDF (Term Frequency Inverse Document Frequency)	15

2.6 Medical Ontologies and Taxonomies	15
2.6.1 UMLS (Unified Medical Language System)	16
2.6.2 SNOMED-CT	16
2.6.3 MedDra (Medical Dictionary for Regulatory Activities)	17
2.7 Techniques for ADR Detection	17
2.7.1 Supervised Text Classification methods	17
2.7.2 Neural Networks	18
2.8 Classification Algorithms	19
2.8.1 Logistic Regression	19
2.8.2 Support Vector Machines	19
2.8.3 Long Short-Term Memory	19
2.8.4 Convolutional Neural Networks	20
2.9 Evaluation	20
2.9.1 Precision	20
2.9.2 Recall	20
2.8.3 Accuracy	21
2.8.4 F1 Measure	21
2.9 Discussion	21
3. Methodology	23
3.1 Proposed System Architecture	24
3.2 Data Collection	25
3.3 Inter – rater Agreement	26
3.4 Data Preprocessing	27
3.5 Feature Selection	29
3.6 Classification Algorithms	32
3.6.1 Logistic Regression	32
3.6.2 Support Vector Machines	33
3.6.3 Convolutional Neural Networks	33

3.6.4 Long Short-Term Memory (LSTM)	35
4. System Evaluation	36
4.1 Baseline Experiments	36
4.2 Preprocessing	37
4.2.1 Stop Words	38
4.2.2 Lemmatization	38
4.2.3 Stemming	39
4.2.4 Stemmed and Lemmatized	39
4.3 Features	40
4.3.1 Word Embeddings	41
4.3.1.1 Glove and Google News Vectors	41
4.3.1.2 Vector Dimensions	42
4.3.1.3 Word Vectors built from the background corpus	42
4.3.1.4 Embedding Layer	44
4.4 Classification Algorithms – Hyper Parameter Tuning	44
4.4.1 CNN	44
4.4.2 LSTM	46
4.4.3 SVM	47
4.4.4 Logistic Regression	47
4.5 Confusion Matrices	48
4.6 Error Analysis	49
4.7 SVM Vs LSTM	50
4.8 Summary	50
5. Conclusion	51
5.1 Future Work	52
References	53

List of Figures		Page
Fig 3.1	The Proposed Architecture	24
Fig 4.1	Confusion matrix for Logistic Regression	48
Fig 4.2	Confusion matrix for SVM	48
Fig 4.3	Confusion matrix for CNN	48

List of Tables

Table 2.1.1	ADE and Non-ADE Examples	07
Table 2.1.2	Distinguishing ADR and Non-ADR	08
Table 3.2	Word2Vec Attributes	31
Table 3.3	CNN Parameters	33
Table 3.4	LSTM Parameters	35
Table 4.1	Baseline experiment results	36
Table 4.2	Initial experiment results	37
Table 4.3	Performance with stop words	38
Table 4.4	Performance with Lemmatization	38
Table 4.5	Performance with Stemming	39
Table 4.6	Performance with Stemming and Lemmatization	39
Table 4.7	All Preprocessing Combined	40
Table 4.8	Feature with Algorithm Results	41
Table 4.9	Glove and Google News Vectors Performance	41
Table 4.10	Glove Word Representation	42
Table 4.11	Most similar words for “cough”	42
Table 4.12	Most similar words for “mucus”	43
Table 4.13	Performance of CNN with Gensims	43
Table 4.14	Performance of Embedding Layer	44
Table 4.15	CNN Conv1D Filter Performance	45

Table 4.16	CNN Dropout Regularization Performance	45
Table 4.17	CNN Activity Regularization Performance	45
Table 4.18	LSTM Spatial Dropout Layer Performance	46
Table 4.19	LSTM Dropout Regularization Performance	46
Table 4.20	LSTM Recurrent Dropout Performance	46
Table 4.21	SVM Inverse of Regularization Strength	47
Table 4.22	LR Inverse of Regularization Strength	47
Table 4.23	Error Analysis	49
Table 4.24	Summary	50

List of Equations

Equation 2.1	Precision Calculation	20
Equation 2.2	Recall Calculation	20
Equation 2.3	Accuracy Calculation	21
Equation 2.4	F1 Calculation	21
Equation 3.1	Kohens – Kappa	27

List of Abbreviations

Abbreviation	Description
ADR	Adverse Drug Reaction
OTC	Over the Counter
FDA	Food and Drug Administration
DoTs	Dose Time and Susceptibility
PV, PHV	Pharmacovigilance
CRM	Customer Relationship Management
URL	Universal Resource Locator
POS	Part of Speech
UMLS	Unified Medical Language System
SVM	Support Vector Machines
ME	Maximum Entropy
MNB	Multinomial Naïve Bayes
CTakes	Clinical Text Analysis and Knowledge Extraction System
jSRE	Java simple relation extraction
NLP	Natural Language Processing
BOW	Bag of Words
TF	Term Frequency
BOAW	Bag of Audio Words
TF-IDF	Term Frequency Inverse Document Frequency
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
ICH	International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human
CRF	Conditional Random Fields
RNN	Recurrent Neural Network

CNN

Convolutional Neural Network

LSTM

Long Short-Term Memory