

**PHARMACOVIGILANCE STUDY OF OPIOID DRUGS ON TWITTER AND
PUBMED USING ARTIFICIAL INTELLIGENCE**

BY

MD. JAMIUR RAHMAN RIFAT

ID: 152-15-5611

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Sheak Rashed Haider Noori, PhD

Associate Professor and Associate Head

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

4 MAY 2019

APPROVAL

This thesis titled “**Pharmacovigilance study of opioid drugs on Twitter and PubMed using artificial intelligence**”, submitted by Md. Jamiur Rahman Rifat, ID No: 152-15-5611 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 4 May 2019.

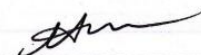
BOARD OF EXAMINERS



Dr. Syed Akhter Hossain
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

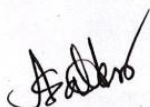
Chairman



Nazmun Nessa Moon
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

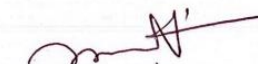
Internal Examiner



Abdus Sattar
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin
Professor

Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

I hereby declare that, this thesis has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head Department of CSE, Daffodil International University**. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:



Md. Jamiur Rahman Rifat
ID: 152 – 15 – 5611
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

I have given my efforts to this thesis. However, it would not have been possible without the kind support and help of many individuals. I would like to express my deepest appreciation to all those who provided me the possibility to complete this report.

At first, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessings which allowed me to complete this thesis successfully.

A special gratitude I give to my supervisor, Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head of CSE department, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my thesis especially in writing this report. His endless patience, scholarly guidance, constant and energetic supervision, constructive criticism, valuable advice have made it possible to complete this thesis.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of my department head, Professor Dr. Syed Akhter Hossain, who provided me with his precious time and kind help to finish this thesis. I also give my deepest thanks to all the faculty members and staff of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Increased usage and under reporting of adverse drug reactions (ADRs) of opioids instigates us to explore some other data sources like Twitter and PubMed. Our paper aims at discovering illegal trafficking of opioids as well as distinguishing tweets from having ADRs or not using binary classifier. We also evaluated the performance of MetaMap in finding ADRs from Twitter and compared the MedDRA encoding system on ADR terms found from tweets and PubMed. We used Latent Dirichlet Allocation (LDA) to find tweets related to illicit sale and used several neural networks for binary classification. It was reported that out of 98 ADRs found from tweets, 50 could be mapped to Lowest Level Terms (LLTs) and 48 to (Preferred Terms) PTs where only 23 LLTs and 15 PTs were reported from PubMed. Among the binary classifier Convolutional Recurrent Neural Network (CRNN) were found to be more promising with .71 F1 score though other models are close to the best one with little margin. Effect of skewness was also monitored in our study. Social media is a good choice for mining pharmacovigilance but during extraction a lot more noise data may come which needs to be avoided.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Output	3
1.5 Report Layout	3
CHAPTER 2: BACKGROUND	4-5
2.1 Introduction	4
2.2 Related Works	4
CHAPTER 3: RESEARCH METHODOLOGY	6-14
3.1 Generating keywords	6
3.2 Data Collection	7
3.3 Manual annotation	7
3.4 Mapping to MedDRA terms	8
3.5 Finding ADRs using MetaMap	9
3.6 Topic modeling	10
3.7 PubMed Exploration	10
3.8 Binary Classification using deep learning	11
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	15-17
4.1 Results	15
4.2 Descriptive Analysis	16

CHAPTER 5: CONCLUSION AND FUTURE WORK	18
5.1 Conclusion	18
5.2 Future Work	18
REFERENCES	19-21
PLAGIARISM REPORT	22

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Some tweets having different spelling of “Morphine”	6
Figure 3.2: Examples of the annotated tweets	8
Figure 3.3: Work flow of our study	8
Figure 3.4: Samples of tweets regarding illegal sale of opioid drugs	10
Figure 3.5: Snippet of XML file format of PubMed data	11
Figure 3.6: Preprocessing example of a tweet	13
Figure 4.1: comparative analysis of different models (50:50 sampling)	16

LIST OF TABLES

TABLE	PAGE NO
Table 3.1: Metaphone encoding output	6
Table 3.2: Example of variant list of some drugs	7
Table 4.1: MedDRA encoding performance	15
Table 4.2: Evaluation metric of different models	15
Table 4.3: Performance of different models on four random tweets	16

CHAPTER 1

INTRODUCTION

1.1 Introduction

Pharmacovigilance is the science and actions pertaining to the identification and prevention of drug related problems and adverse events [1]. Adverse drug reactions (ADRs) is the partial unsuccessful and harmful contribution of a drug on a human body. Pharmacovigilance is the study of drug safety monitoring which has the potential to find ADRs that were previously unknown [2]. It uses the feedback of patients to improve the medication system and creates faith on medicinal products among the general mass [3].

Medical Dictionary for Regulatory Activities (MedDRA) which is used as a standard lexicon for classifying adverse event by pharmaceutical industry and health regulatory authorities for reducing ambiguity among themselves during the regulatory process, from pre-marketing to post-marketing activities. In MedDRA the terminologies are classified into five hierarchies which provides options for retrieving data by specific or broad grouping [4]. We have used MedDRA for mapping the user specified ADR terms to Unified Medical Language System (UMLS) Concept Unique Identifier (CUI). For a single CUI we tried to identify its associated LLTs and PTs. Nikfarjame et al. [5] manually mapped patients defined ADR terms to UMLS CUI by the medical practitioners but we employed UMLS REST Application Programming Interface (API) to fasten the mapping process. MetaMap is a program which can be used to find clinical text from biomedical body using NLP or Computational linguistics [6]. We have used MetaMap to find ADR terms form Twitter and PubMed.

To identify whether a tweet contains any ADR related terms or not, we built a binary classifier using deep learning methods. Ginn et al [7] used machine learning approaches like naive bayes and SVM for binary classification on tweets. But being inspired from the work of Huynh et al [8] we used deep learning for avoiding manual feature engineering.

Finally, we would use PubMed, a free search engine for accessing biomedical literature of MEDLINE database containing more than 28 million citations as a source of structured data for finding ADRs [9]. The overall workflow of our study has been demonstrated in figure 1.

1.2 Motivation

A review study of Sarkar et al [10] had shown that ADRs accrue an annual cost of 75 billion dollars in hospital related activities. Opioids are a class of drugs used for pain management and anesthesia. Opioids can cause a significant burden on health care systems and have become a more remarkable contributor to health care costs due to ADRs in recent decades. It has been estimated that the monetary toll of prescription opioid abuses and overdoses in the United States was 78 billion dollars in 2013 with only 3.6% accounting for legitimate medical treatments [11]. As Opioid use has tripled from 1991-2011 thus post market surveillance of opioid related issues needs to be increased [12]. Though Government agencies have developed tools such as US FDA's MedWatch program or the UK MHRA's Yellow Card Scheme to collect ADRs directly from patients and medical practitioners using these systems. However, these reporting platforms can only capture 10% of opioid related adverse events [13]. Due to these limitations, researchers have turned to social media platforms such as Twitter, DailyStrength, PatientsLikeMe for mining ADRs. DailyStrength and PatientsLikeMe are health specific networks and are more useful to health researchers. Twitter is a great source of data with 500 million tweets sent per day [14]. In this study we will use Twitter and PubMed citations for pharmacovigilance research. Identifying ADRs in tweets is very challenging because:

1. Tweets are usually written colloquially without following proper grammar and correct spelling.
2. People have no knowledge of medical terminology regarding their drug effect.

In addition to ADRs, the illegal trade of opioids spreads drug addiction. It is possible to identify illegal trading information from Twitter using an unsupervised model called topic modeling. Topic modeling is a method of finding the subject in a document and can be used as classifying text of bulk size where the supervised methods are much harder and impossible to operate.

1.3 Rationale of the Study

There are much research that were carried on pharmacovigilance mining of general drugs. But for the opioid drugs such exploration is so scarce. Also, opioid drugs bear much challenges that other drugs don't have, like illegal trafficking.

1.4 Outcome

This research work aims at

1. Classifying tweets in binary format to know whether a tweet contains ADR information or not.
2. Identifying illegal sale of opioid drugs from Twitter.
3. Comparing the performance of MetaMap and MedDRA encoding system.

1.5 Report Layout

In this chapter we have discussed about the introduction of pharmacovigilance study, motivation, rationale of the study and the outcome of the thesis. Later followed by the report layout.

In chapter 2, we will discuss about the background of our research topic.

In chapter 3, we will discuss about the methodologies employed in our study.

In chapter 4, we will discuss about the obtained results and discussion.

In chapter 5, we will discuss about the conclusion and future work.

CHAPTER 2

BACKGROUND

1.1 Introduction

Previously several researches have been conducted to find the adverse effect of drugs from social media or national databases as the inverse drug effect can be alarming for the national health care system. In such cases, the work of Glowacki et al. [15] showed that the public concern about the opioid misuse can bring havoc. To tackle those situations much research has been done, though are not sufficient.

1.2 Related Works

In a paper of Henriksson et al. [16], they have used electronic health records (EHR) to identify adverse drug events (ADEs) as the in practice discretionary reporting system was so naïve. They used distributional semantic representation which is a type of unsupervised learning for finding named entities such as disorders, symptoms, drugs etc. Sarkar and Gonzalez [17] classified ADR texts from three annotated datasets. One of which was from clinical reports and other two were from social networks. This study also claimed that the usage of NLP techniques will be beneficial for creating high quality feature set.

The present trending and usefulness of using social media data has been come out from the review research output of Sarker et al. [18] where out of twenty two projects only six had the annotation corpus. That insights us that the availability of annotated corpus is still very less and unsupervised learning can be more efficient in this case.

In the study of Ginn et al. [19], 10822 tweets were annotated manually. This paper gave us the outline of annotating the corpus. The corpus was annotated for binary classification with kappa value .69. The corpus was trained using Naïve Bayes (NB) and Support Vector Machine (SVM).

The work of Sarker et al. [20] tried to find that “is it possible to use Twitter for finding medical prescription abuse”. They found that social media is a viable source of finding adverse prescription uses.

Other than social media, online health forums can also be used for mining pharmacovigilance which has been found in the study of Chee et al. [21]. Texts were classified using ensemble classifier such as boosting, bagging.

Zorzi et al. [22] developed a language independent NLP algorithm named “MagiCoder” for encoding user specified terms to MedDRA terms.

The guidelines for preparing the keyword List for acquiring tweets were found from the work of Pimpalkhute et al. [23], where they have used phonetic spelling variants, as in Twitter the users are not much concerned about the spelling of a drug name. The google custom search engine was used in their work for counting the number of usage of that word.

Huynh et al [8] harnessed different types of neural network for classifying tweets related to ADRs in binary format.

A detailed work on pharmacovigilance had been demonstrated in the paper of Nikfarjam et al. [24]. They not only classified the text in binary format but also tried to find named entities such as name of the drug, span of ADR terms, span of Indication terms from the text. The performance of their model was differentiated between two types of data sources. One is Twitter and the other one is DailyStrength (DS). The machine learning methods used by them were Conditional Random Fields (CRF), clustering and word embeddings. The model worked best on DS datasets.

Limsopatham and Collier [25] tried to encode the user specified ADR terms with the medical ontologies using deep neural network. But we used MedDRA encoding scheme in our work.

Katsuki et al [26] used Support Vector Machine (SVM) and protocol of content coding for finding illegal sale data. But using supervised learning for finding those data from a colossal of data is troublesome. Mackey et al [27] used Biterm Topic Model for illicit drug sale identification but Latent Dirichlet Allocation (LDA) was employed in our work.

Norbutas [28] explored crypto market for finding the structure of illegal trading of opioid drugs. He used Exponential Random Graph Models (ERGM) to accomplish his goal.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Generating keywords

We have downloaded tweets using a total of 167 keywords using the Stream API of Twitter. Primarily generic names of 17 opioid drugs such as fentanyl, codeine and 11 opioid class names such as opioid medication, analgesics, pain medication were chosen based on the recommendation of a pharmacology expert. At first using one edit distance algorithm we created all the possible variants of a name. The drug name “morphine” will have its associated variants list like morphin, morfine and moaphine. It is assumed that phonetic misspelling occurs more frequently than other sorts of errors. Therefore, we kept those name in the variant list which has similar phonetic representation with respect to the original name. We used python metaphone library to get the pronunciation words. The metaphone encoding output has been represented in table 3.1. Then from the remaining variant list we choose the final set of variant list manually by using the Google hit counts. The Google Custom Search Engine API [29] was used to calculate the number of google hits. Figure 3.1 illustrates examples of hits from our methods:

“I had to raise slow down & rest cause of stronger pains this lasr week. More #Morphin I have already had several operations of the vertebral column and I'm going to have it others in the coming months.”
“GIVE ME SOME MORFINE.”

Figure 3.1 Some tweets having different spelling of “Morphine”

Table 3.1 Metaphone encoding output

Variant Words	Metaphone encoding
morphine	MRFN
morfine	MRFN
moaphine	MFN
morphin	MRFN

After using the above mentioned procedure we got the final variant list some of which has been endorsed in table 3.2.

3.2 Data Collection

We have collected data from Twitter using the Twitter Stream API from (15 February 2018-24 April 2018). A total of 166723 tweets were downloaded. Out of which 68204 tweets were taken primarily for annotation. All the 166723 tweets were taken for topic modeling. As tweets are very unstructured in nature and contain a lot of unicode characters, we need to preprocess the raw form of the data. Manually annotated data was first trimmed based on language. We kept the data which belongs to English language. Later, removed retweets based on condition.

data starts with “RT” → remove tweet

Using another rule, we have removed the tweets containing external web links.

data contains “https” → remove tweet

Table 3.2 Example of variant list of some drugs

Original Drug Names	Example of variants
Tramadol	tramadoll, tramadel
morphine	morfine, morphin
codeine	codene, codine

Finally, we removed duplicated tweets and unicode characters from tweets. Thus we got 4633 tweets for manual annotation. These data will also be used by the binary classifier.

In case of topic modeling, we kept the retweets and tweets containing external links. However, before removing duplicated tweets we also removed unicode characters, ‘@’ and ‘#’ characters and punctuations and numerical values. Finally, we have 60140 tweets to feed to our topic modeling algorithm.

3.3 Manual annotation

After preprocessing we got a total of 4633 for manual annotation. We annotated the body in two ways. One for the binary classification to determine whether a tweet is

containing an ADR information or not. We have classified the tweets containing ADRs as 1 and not containing ADRs as 0. A total of 98 tweets are classified as 1 that means those tweets contain signals of having ADR related data. Tweets classified as 1 are taken for annotation in details. We have annotated the corpus in details mentioning the drug name, span of expression mentioning either ADR or indication. We consider, ADR is regarded as the inverse effect of a drug for its consumption. while indication is for which a drug is prescribed for or taken during illness. Finally, we got 89 tweets containing ADRs and 9 tweets bearing indications. Figure 3.2 delineates some of the annotated tweets.

“my friend is a fucking oxy (drug name) fiend (idk why because that shits expensive) and her nose bleeds (ADR) lots too”
“Misplaced a vial of morphine (drug name) today. Turns out it causes quite a commotion...(ADR)”
“Have ambiguous genitalia?You probably have Toxic Shock Syndrome (Indication).This causes a total eclipse of the heart.Try Methadone (drug name)”

Figure 3.2 Examples of the annotated tweets

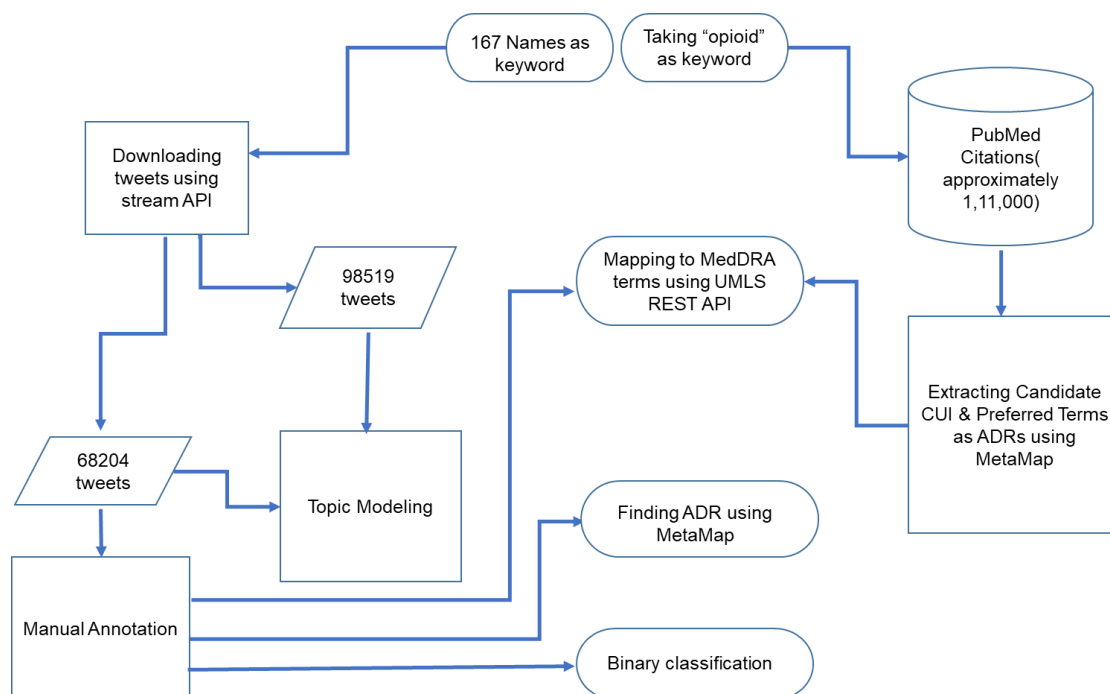


Figure 3.3 Work flow of our study

3.4 Mapping to MedDRA terms

MedDRA is the abbreviated form of Medical Dictionary for Regulatory Activities which is used as a standard lexicon for classifying adverse event by pharmaceutical

industry and health regulatory authorities for reducing ambiguity among themselves. In MedDRA the terminologies are classified into five hierarchies. LLTs is at the most specific level containing more than 70,000 terms. The next level is PTs which has at least one LLT or lexical variants. Related PTs are grouped to HLT terms and subsequently same HLTs are incorporated to HGLTs. Lastly the HGLTs are assembled to most general terms SOC based on etiology, manifestation site or purpose [4]. As MedDRA is a standard dictionary for medical terminology we can get insight into how people report ADRs in their own terms without having knowledge about medical jargon. We have considered a successful mapping of the ADR terms if we could get the corresponding CUI value of that term using the API. To retrieve CUI by searching with ADR terms, the path value “/search/{version}” was used with the base URI. The parameters value for “searchType” and “sabs” were chosen “exact” and “MDR” respectively. Later, to find the LLT and PT terms of the CUI values found from the above query we have used the path value “/content/{version}/CUI/{CUI}/atoms” with the base URI [30]. At last we recorded that 51.02% annotated ADR terms were successfully mapped to MedDRA terms. The result is compiled and illustrated in table 4.1.

3.5 Finding ADRs using MetaMap

MetaMap is tool for discovering UMLS terms from texts using computational linguistic procedure. It seems like an act of information retrieval. The tool can be used in three way. First one is using MetaMap interactively, that means directly getting the output by writing some texts on the interactive web platform. Second one is Batch Metamap, in which users uploads file of texts and the MetaMap generates an output in the form of json, plain text etc. The last one is using the web API [6].

In this study we wanted to find the number of terms related to ADRs (including indications) from our annotated tweets body using batch MetaMap. Then compared the result with the manual identification process. Batch MetaMap returned a xml file where each token was given by a semantic type value. The tokens under a certain semantic type were considered as ADRs. The semantic types are as follows: injury or poisoning, pathologic function cell or molecular dysfunction, disease or syndrome, experimental model of disease, finding, mental or behavioral dysfunction neoplastic process, signs or symptoms, mental process [5]. We have got 137 terms as ADRs and extra 39 terms as ADRs.

3.6 Topic modeling

All 60140 preprocessed tweets were fed in our algorithm. As LDA is very promising in this era so we used the gensim implementation of LDA. We have selected the number of topic as 10 and chosen our desired topic which was represented by the words like buy, spend, price. Then for each tweet we build a topic distribution. The topic which got the maximum distribution value is regarded as the representational topic of that tweet. Thus, if the representational topic matches with our target topic then we have marked that tweet as bearing a signal of illicit drug marketing and sales. Finally, we got 2880 tweets as such though there are some tweets which are not related to illegal sale. Figure 3.4 gave us some example of such tweets. But this method shrank the size of manual monitoring and helped us to get some valid examples. Some instances are given below. We have also kept the associated external web links in a separate column for further forensic examination of the websites.

“Buy Hydrocodone online Street value Without Prescription. Secure payment overnight freeship delivery in usa.”

“Buy Tramadol Online Video Looking for a tramadol? Not a problem! Buy tramadol online ==>”

“take DayQuail or Nightquail”

“I recently bought your Hydro Skin Lip Therapy product.. But I spend more time trying to get it out the nozzle”

“Tulsans and Oklahomans who live nearby spend between \$18.7 and nearly \$21 million per year on heroin. That's money”

Figure 3.4 Samples of tweets regarding illegal sale of opioid drugs

3.7 PubMed Exploration

Using “Opioid” as keyword approximately 1,11,000 biomedical literatures from PubMed with abstract text and abstract title in xml format were downloaded. A snippet of the data has been figured out in Figure 3.5. There were a lot of tagset in the file but we have extracted the PMID and AbstractText tagset in our study. A snippet of our file

has been illustrated in figure1. As the data set was huge so we took 8.11% of the data that is 9005 citations. Each Citations is uniquely defined with a PMID. To get ADR terms from the abstract texts we used MetaMap and considered previously mentioned [5] semantic types to be effectively tagged as ADR. By doing so we got a total of 57168 tokens as ADR. For a single PMID we got a couple of Preferred terms tagged as ADR. But it is quite challenging to find the exact PMID from the returned xml file by MetaMap. To identify the true PMID we have checked whether the PMID is a digit or not and is equal to the PMID returned from the PubMed. To specifically mapping to MedDRA, our further work will expand to find LLTs and PT terms from those tokens. In table 3 we demonstrated a comparison of MedDRA coding of Twitter and PubMed. For our access limitations we took 100 random ADR terms to feed into our mapping procedure.

```

<PubmedArticle>
  <MedlineCitation Status="Publisher" Owner="NLM">
    <PMID Version="1">29700475</PMID>
    <DateRevised>
      <Year>2018</Year>
      <Month>04</Month>
      <Day>27</Day>
    </DateRevised>
    <Article PubModel="Print-Electronic">
      <Journal>
        <ISSN IssnType="Electronic">1546-1718</ISSN>
        <JournalIssue CitedMedium="Internet">
          <PubDate>
            <Year>2018</Year>
            <Month>Apr</Month>
            <Day>26</Day>
          </PubDate>
        </JournalIssue>
        <Title>Nature genetics</Title>
        <ISOAbbreviation>Nat. Genet.</ISOAbbreviation>
      </Journal>
      <ArticleTitle>Genome-wide association analyses identify 44 risk variants and refi
      <ELocationID EIdType="doi" ValidYN="Y">10.1038/s41588-018-0090-3</ELocationID>
      <Abstract>
        <AbstractText>Major depressive disorder (MDD) is a common illness accompanied
      </Abstract>
    </Article>
  </MedlineCitation>
</PubmedArticle>

```

Figure 3.5 Snippet of XML file format of PubMed data

3.8 Binary Classification using deep learning

In machine learning, classification is a task of assigning a new instance to a preordained category based on training set of data whose true class is already known [31]. Therefore, in an abundance of twitter data where people used to write about diverse topics from sports to science and their day to day life activities, finding an ADR related data will

be sparse and it will be almost impossible for human to find those ADR signal data manually by reading between lines. So, building a classification model will eliminate the manual monitoring system. Several machine learning approaches were used previously like Support Vector Machine (SVM) [32], logistic regression [33]. Introduction of deep learning is so innovative currently as it does not require explicit feature engineering for natural language processing task. We have used Convolution Neural Network (CNN)[8] and Recurrent Neural Network (RNN)[8] with some of their variants to see which one works better. We also evaluated the sampling effect because our dataset was highly skewed. We have got only 98 tweets as true positives out of 4633 tweets. We resampled the data in two ways. At first taking equal number of class “1” and class ”0” data (50:50 ratio). Another one was constituted taking 6:94 ratio where 98 tweets of class containing ADR were shuffled with 1400 tweets of class not containing any ADR information data. The evaluation of different models are given in table 4.2.

3.8.1 Preprocessing

First of all, we converted all letters to small and tokenized the word using NLTK tokenizer, then using regex removed punctuation marks and digits. We stemmed the word to its root form using WordNetLemmatizer. Tweets contain a lot of spelling errors hence it would be better to automatically correct a misspelled word. But for the computational complexity we have not deployed this method. The whole procedure has been illustrated with an example in figure 3.6. Each token has been separated using “/”.

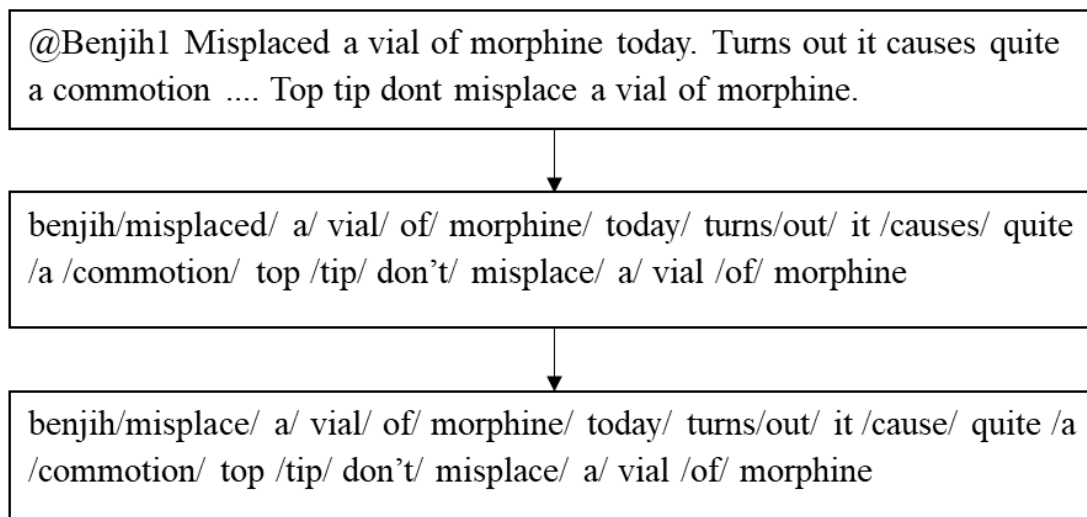


Figure 3.6 Preprocessing example of a tweet.

3.8.2 Word Embeddings

In natural language processing when using deep learning word embeddings plays a role to identify the meaning of a word. Each word is uniquely defined by a vector of real numbers. The shape is about $[m,v]$ where m = number of words in the vocabulary and v = number of representational states. The more the number of representational states, the more the accuracy of the model will be and the computational cost will be greater also. Building a quality embeddings requires huge amount of data and time to run. Therefore, we have used pretrained Global Vector of stanford [34] which is of 50 dimensions.

3.8.3 Models

CNN, Convolutional Recurrent Neural Network (CRNN), RNN, Recurrent Convolutional Neural Network (RCNN) were implemented using the sequential model of keras and Convolutional Neural Network with Attention (CNNA) with the functional API. “Relu” activation function was applied in the hidden layers. For optimization deployed “adam” optimizer in all models.

CNN

The model started with an embedding layer. A dropout value of .25 was used. In the convolution layer we have used a filter size of 32. In , used a max pooling layer of pool size 2. Then the input nodes were flatten. We used a hidden layer of 250 nodes. In the output layer used a single node and sigmoid was used as an activation function in the output node.

CRNN

The first layer was an embedding layer. Then the output was convolved using a filter size 32. Max pooling was operated using a pool size 2. Then instead of flattening we used a Gated Recurrent Unit (GRU) layer of 300 nodes. At last, sigmoid function was used to generate the output.

RNN

This model also started with an embedding layer. We used GRU layer of 300 nodes which is a variant of RNN. Then used a dropout value of 0.5. The output layer is a single node with “sigmoid” activation function.

RCNN

Embedding layer was the starting step as usual. A basic RNN layer comprising 300 nodes were then used. Then a convolutional layer, max pooling layer and flattening were performed respectively. A dropout value of 0.50 was used. Lastly, the final output was a single noded layer with sigmoid activation function.

CNNA

This model was initiated with an embedding layer. Then applied a dropout value of 0.25. Using a filter value of 32 the output matrix was convolved. After that we incorporated the attention mechanism. In the attention layer “softmax” activation function was used. The nodes were flatten and a single final output node was used with a sigmoid activation function.

Our models give a probability value of the class. If the probability value is more than .50 then we classified that tweets as class “1” that means it contains ADR related information and vice versa.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results

Table 4.1 MedDRA encoding performance

Study domain	Total number of ADR terms	Failed to map	LLTs found	PTs found
Twitter	98	48	50	48
PubMed	100	77	23	15

The table 4.1 demonstrates a surprising result where it was believed that MedDRA is good at mapping structured data sources like PubMed but we got that the encoding platform works much better in social media sites like Twitter. In PubMed we found less number of PT terms compared to the LLT terms found.

We have evaluated our different methods using (recall, precision and f1 score) by splitting our dataset into 70:30 ratio where 70% data were used for training and 30% data for testing. The performance of the models are close to each other where CRNN performs better than others. We have also observed that skewness could highly affect our model performance. In table 4.3 we also noticed the predicted result of our models on four tweets.

Table 4.2 Evaluation metric of different models where bests are marked with bold

Method	50:50 sampling			6:94 sampling		
	Recall	Precision	F1	Recall	Precision	F1
CNN	.71	.68	.69	.10	.25	.14
CRNN	.71	.71	.71	.14	.14	.14
RCNN	.57	.66	.61	.10	.20	.13
RNN	.61	.72	.66	.14	.13	.13
CNNA	.72	.65	.68	.05	.12	.07

The results obtained from various models doesn't fluctuate a lot but overall CRNN came up victorious with f1 score .71. The figure 4.1 demonstrates the comparative supremacy of each model.

Table 4.3 Performance of different models on four random tweets

Tweets	True class	Predicted Class				
		CNN	CRNN	RCNN	RNN	CNNA
<i>"Last time was paracetamol. It goes: paracetamol, codine, morphine in-terms of painkillers"</i>	0	1	0	0	0	0
<i>"Have trouble focusing? You probably have ADHD.This causes denial.Try Methadone"</i>	1	1	0	0	0	1
<i>"This squirrel needs Vicodin!"</i>	0	0	0	0	0	0
<i>"Have hardening of the nipples? You probably have PTSD.This causes a catatonic state.Try Xanax"</i>	1	1	1	1	1	1

4.2 Descriptive Analysis

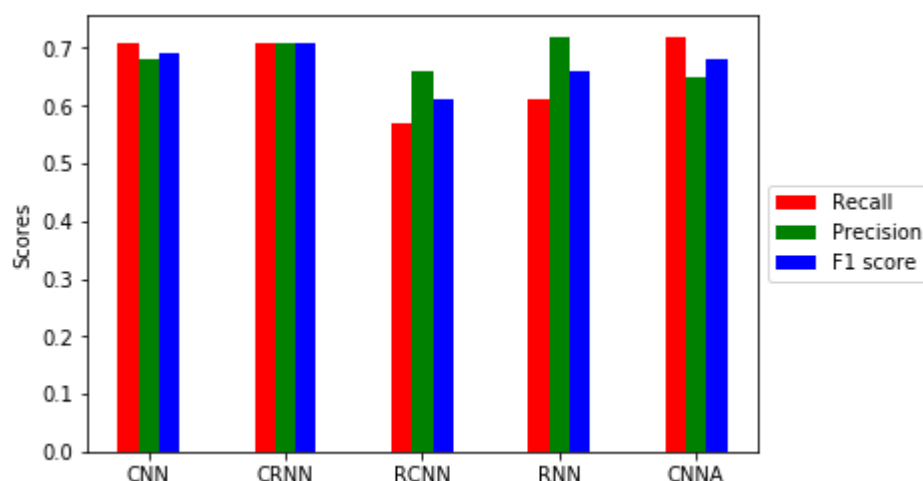


Figure 4.1 Comparative analysis of different models (50:50 sampling)

It was observed that people are less interested to share any negative impact on their body of any drug in social media. When it is opioids the result is much less as expected. On the contrary it was seen that the number of tweets regarding illegal sale of opioid drug is quite alarming. For pharmacovigilance other social network sites like DailyStrength, PatientsLikeMe or any other drug forum will be more convincing than Twitter. But Twitter provides free access for anybody. Facebook can also be great source of data for such study.

Though we have collected data about 17 drug names. But after manual annotation in our corpus it was noticed that codeine, fentanyl, tylenol were dominated in ADR report. Another think that was noted in our report that codeine is a popular pain killer and

sometimes it is termed as cody. But this keyword generates a lot of noise data which is not our case of study.

As word embeddings helps to understand the meaning of different unknown words and in biomedical text research there could be many medical terminologies, so the choice of word vector representation can contribute to the success of any deep learning methodology. Though we have used GloVe but other methods like word2vec and FastText could be effective also.

Albeit we used supervised classification methods but unsupervised models like autoencoder and restricted boltzman machine can be used for binary classification. In an environment where it is difficult to get labeled data, unsupervised model can be a best choice [35].

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In our study we have assembled different techniques those were explored individually before to analyses pharmacovigilance from twitter and PubMed for opioid drugs. It is a specialized class of drugs which got several misuses than any other drugs. Illegal distribution of those drugs can rise in drug addiction in the society. Our system can find those information from a bulk of data size. We also justified the performance of MetaMap on both PubMed data and twitter. The binary classifier built depicts that the availability of quality data is much more appreciable than the choice of algorithm as the performance is same for different models. One of the drawback of our study is the lack of quality data which can be eradicate by using some other health related social media sites. But if anybody wants to use twitter or Facebook then they need to define the keyword list as precise so that noise data couldn't jumble up.

5.2 Future Work

Here we have just classified tweets to know whether it contains ADR terms or not but we will look forward to extracting ADR terms from texts. We would also like to use clustering to identify illegal marketing tweets.

REFERENCES

- [1] O'Connor, Karen, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L. Smith, and Graciela Gonzalez. "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions." In *AMIA annual symposium proceedings*, vol. 2014, p. 924. American Medical Informatics Association, 2014.
- [2] Shaw, Debbie, Ladds Graeme, Duez Pierre, Williamson Elizabeth, and Chan Kelvin. "Pharmacovigilance of herbal medicine." *Journal of ethnopharmacology* 140, no. 3 (2012): 513-518.
- [3] World Health Organization. "The importance of pharmacovigilance." (2002).
- [4] MedDRA (2018) <https://www.meddra.org/>. Accessed 6 June 2018.
- [5] Nikfarjam, Azadeh, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features." *Journal of the American Medical Informatics Association* 22, no. 3 (2015): 671-681.
- [6] MetaMap (2018) A Tool For Recognizing UMLS Concepts in Text. <https://metamap.nlm.nih.gov/>. Accessed 8 June 2018.
- [7] Ginn, Rachel, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. "Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark." In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*. 2014.
- [8] Huynh, Trung, Yulan He, Alistair Willis, and Stefan Ruger. "Adverse drug reaction classification with deep neural networks." *Coling*, 2016.
- [9] PubMed (2018). <https://www.ncbi.nlm.nih.gov/pubmed/>. Accessed 20 April 2018.
- [10] Sarker, Abeed, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. "Utilizing social media data for pharmacovigilance: a review." *Journal of biomedical informatics* 54 (2015): 202-212.
- [11] NIH(2018) Medications to Treat Opioid Use Disorder. <https://www.drugabuse.gov/publications/research-reports/medications-to-treat-opioid-addiction/how-much-does-opioid-treatment-cost>. Accessed 13 July 2018
- [12] NIH (2018) Prescription Opioids and Heroin. <https://www.drugabuse.gov/publications/research-reports/relationship-between-prescription-drug-abuse-heroin-use/increased-drug-availability-associated-increased-use-overdose>. Accessed 13 July 2018
- [13] Inman, William, and Gillian Pearce. "Prescriber profile and post-marketing surveillance." *The Lancet* 342, no. 8872 (1993): 658-661.
- [14] Omnicore (2018) Twitter by the Numbers: Stats, Demographics & Fun Facts. <https://www.omnicoreagency.com/twitter-statistics/>. Accessed 20 July 2018
- [15] Glowacki, Elizabeth M., Joseph B. Glowacki, and Gary B. Wilcox. "A text-mining analysis of the public's reactions to the opioid crisis." *Substance abuse* (2017): 1-5.

- [16] Henriksson, Aron, Maria Kvist, Hercules Dalianis, and Martin Duneld. "Identifying adverse drug event information in clinical notes with distributional semantic representations of context." *Journal of biomedical informatics* 57 (2015): 333-349.
- [17] Sarker, Abeed, and Graciela Gonzalez. "Portable automatic text classification for adverse drug reaction detection via multi-corpus training." *Journal of biomedical informatics* 53 (2015): 196-207.
- [18] Sarker, Abeed, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. "Utilizing social media data for pharmacovigilance: a review." *Journal of biomedical informatics* 54 (2015): 202-212.
- [19] Ginn, Rachel, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. "Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark." In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*. 2014.
- [20] Sarker, Abeed, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter." *Drug safety* 39, no. 3 (2016): 231-240.
- [21] Chee, Brant W., Richard Berlin, and Bruce Schatz. "Predicting adverse drug events from personal health messages." In *AMIA Annual Symposium Proceedings*, vol. 2011, p. 217. American Medical Informatics Association, 2011.
- [22] Zorzi, Margherita, Carlo Combi, Riccardo Lora, Marco Pagliarini, and Ugo Moretti. "Automagically encoding adverse drug reactions in MedDRA." In *Healthcare Informatics (ICHI), 2015 International Conference on*, pp. 90-99. IEEE, 2015.
- [23] Pimpalkhute, Pranoti, Apurv Patki, Azadeh Nikfarjam, and Graciela Gonzalez. "Phonetic spelling filter for keyword selection in drug mention mining from social media." *AMIA Summits on Translational Science Proceedings 2014* (2014): 90.
- [24] Nikfarjam, Azadeh, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features." *Journal of the American Medical Informatics Association* 22, no. 3 (2015): 671-681.
- [25] Limsopatham, Nut, and Nigel Henry Collier. "Normalising medical concepts in social media texts by learning semantic representation." (2016).
- [26] Katsuki, Takeo, Tim Ken Mackey, and Raphael Cuomo. "Establishing a link between prescription drug abuse and illicit online pharmacies: analysis of Twitter data." *Journal of medical Internet research* 17, no. 12 (2015).
- [27] Mackey, Tim K., Janani Kalyanam, Takeo Katsuki, and Gert Lanckriet. "Twitter-based detection of illegal online sale of prescription opioid." *American journal of public health* 107, no. 12 (2017): 1910-1915.

- [28] Norbutas, Lukas. "Offline constraints in online drug marketplaces: An exploratory analysis of a cryptomarket trade network." *International Journal of Drug Policy* 56 (2018): 92-100.
- [29] Google Custom Search (2018) <https://developers.google.com/custom-search/>. Accessed 13 May 2018
- [30] NIH (2018) UMLS API Technical Documentation. <https://documentation.uts.nlm.nih.gov/rest/home.html>. Accessed 12 May 2018
- [31] Wikipedia (2018) Statistical classification https://en.wikipedia.org/wiki/Statistical_classification. Accessed 8 August 2018
- [32] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20, no. 3 (1995): 273-297.
- [33] Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. "An introduction to logistic regression analysis and reporting." *The journal of educational research* 96, no. 1 (2002): 3-14.
- [34] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.
- [35] Pumsirirat, Apan, and Liu Yan. "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine." *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 9, no. 1 (2018): 18-25.

PLAGIARISM REPORT

PHARMACOVIGILANCE STUDY OF OPIOID DRUGS ON TWITTER AND PUBMED USING ARTIFICIAL INTELLIGENCE

ORIGINALITY REPORT

25%	21%	14%	21%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	9%
2	ceur-ws.org Internet Source	1%
3	Submitted to Fachhochschule Salzburg GmbH Student Paper	1%
4	Submitted to University of Nottingham Student Paper	1%
5	Juan Antonio Lossio-Ventura, Jiang Bian. "An inside look at the Opioid Crisis over Twitter", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018 Publication	1%
6	Submitted to University of Melbourne Student Paper	1%
7	academic.oup.com Internet Source	1%

anthology.aciweb.org