S	
E1	
E2	
For Office Use Only	

Masters Project Final Report (MCS) 2019

Project Title	**Classifying Drug Adverse Events using Social Media Data**
Student Name	C S Gunarathna
Registration No. **& Index No.**	2015/MCS/032 15440322
Supervisor's Name	Dr. Manjusri Wickramarathne

For Office Use ONLY

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: C S Gunarathna

Registration Number: 2015/MCS/032

Index Number: 15440322

C S Gunarathna

Signature: Date: 16/11/2020

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name:

Signature: Date:

Classifying Drug Adverse Events using Social Media Data

A dissertation submitted for the Degree of Master of Computer Science

C S Gunarathna

University of Colombo School of Computing

2020

Acknowledgement

I would like to express profound gratitude to my advisor, Dr. Manjisri Wickramarathne, for his invaluable support by providing relevant knowledge, materials, advice, supervision and useful suggestions throughout this research work. His 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 Mr. Bhattacharya Partha (Director Global Delivery - IQVIA) and Mr. Upeka Karunapathi (Team Lead Software Engineering – IQVIA Solutions Pvt. Ltd). Further I would like to thank all my colleagues for their help on finding relevant research material, sharing knowledge and experience and for their encouragement.

I am as ever, especially indebted to my parents & 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.

Table of Content

Table of Figures

Table List

Abstract

On time recognition of medication unfriendly occasions has been a basic issue for the pharmaceutical business, the conventional route was to recognize them at clinical preliminaries and after the medication arrives at the market gather the patient gripes from specialists. Be that as it may, this procedure devours time and has the danger of missing significant medication unfriendly responses.

On time location of medication unfriendly occasions has been a basic issue for the pharmaceutical business, the customary path was to identify them at clinical preliminaries and after the medication arrives at the market gather the patient whines from specialists. In any case, this procedure expends time and has the danger of missing significant medication unfavorable responses. This research focuses on a Machine Learning approach to screening continuously from social media data such as twitter about drug adverse reactions and classified them from these contents.

Chapter One

Introduction

1.1 Background

A drug, medical appliance or medication is used to either diagnose, cure, prevent or treat a disease, it is one of the most important parts in the field of medicine and the science behind developing these drugs is known as pharmacology [1]. Drugs can be mainly classified into two, namely:

- Prescription Drugs – A legal prescription from licensed medical practitioner needed for getting in a pharmacy and intended to be used only one person
- OTC (Over the counter drugs) – These drugs can be obtaining any grocery or pharmacy without any prescription

After years of research with deeper expertise getting a drug launch right is critical for its overall success. This associate with huge cost investment and had to go with preclinical testing, investigations, clinical studies and new drug application filling,etc. It should validated through many rules and regulations enforced by region, government and related authorities. New drug launches face more intense competition today than they faced a decade ago.

There are two main methods of testing a drug, namely clinical studies and clinical trials. Clinical studies also known as observational studies observe people in normal settings after grouping people into different categories by researchers and then compare changes over time. Clinical trials are aimed at evaluating medical, surgical or behavioral intervention of people taking part, they often check on new drugs or devices and assess whether the new treatment has less side effects and more effectiveness compared to the standard treatment [2].

After years of careful laboratory research experiments in animals and human cells, researchers send the data to the Food and Drug Administration (FDA), which is the government body governing the drug manufacturing in US should approve the tests for approval before testing in humans.

1.2 Problem Definition

Although the physicians are supposed to report possible drug ADRs it is rare that a patient comes back and reports it to the physician, hence most of the drug ADRs go unreported which can be very dangerous. Finding these unreported adverse events is a major challenge for the pharmaceutical companies, but lately people tend to post these extensively on social media [53]. Thus, pharmaceutical companies must analyze through unstructured data like Facebook comments, tweets, call center records of health insurance companies and CRM (Customer Relationship Management) systems of pharmaceutical company sales representatives etc.

The data that is available in social media is highly important as it provides an interaction with the patient in a personal level, but this data is massive and manual evaluation to track down ADRs is a very labor-intensive task. Machine-learning based approaches can be used to analyze the data and flag the data that may potentially contain adverse events which can be further analyzed by professionals to confirm, this way it will reduce the human error that happens when going through that amount of data and it will also reduce the amount of expert labor required. The FDA states their regulation as an ADR should be reported within 24 hours of occurring [54], but with these manual methods achieving this timeline is impossible.

As a solution companies like IQVIA [57] has come up with products like AE Tracker which scans the social media and forums and blogs, but they use a medical ontology to detect adverse events from social media, this approach has a very high false positive rate as well as the ontology should be kept up to date.

There is a considerable research effort in classifying whether a social media content is an adverse event or not but the automatic extraction of this adverse effect from the social media content is not much done. Also, mostly suboptimal ways such as Support Vector Machines & Naive Bayes has been commonly used to automatically learn classifying drug ADR related data from social media data. There have been very few efforts on employing modern methods such as Deep Learning & Word embeddings for mining drug ADRs from social media data, this fact is well demonstrated in the literature review section which looks at similar research. Also, research has considered only popular social media platforms such as Twitter and Facebook, but more valuable information exists in other social media platforms [56].

Social media data constantly updates, hence the learnt model should be constantly updated too, for this we need to employ online learning, but the research done in this area so far has not used online learning for ADR detection.

Drugs and Clinical Trials

A drug or medication is used to either diagnose, cure, prevent or treat a disease, it is one of the most important parts in the field of medicine and the science behind developing these drugs is known as pharmacology. Drugs can be mainly classified into two, namely:

❖ Prescription Drugs – A drug dispensed by a pharmacy only with a prescription from a qualified physician or a nurse.

❖ OTC (Over the counter drugs) – The drugs that can be obtained from groceries or pharmacies without any prescription from a doctor, these are usually very basic drugs such as Aspirin.

Drug Adverse Reactions

Detecting adverse reactions that occur from drugs have become a very important task in today's pharmaceutical and healthcare industry, rules and regulations are being brought up governing the field of adverse events that occur from drugs, specifically by the FDA(U.S. Food and Drug Administration), in the United States Pharmaceutical companies are responsible for documenting and reporting adverse events related to the drugs they manufacture.

A drug adverse reaction is an unintended and harmful event that occurs due to the usage of the drug. Any drug will have the potential of causing an ADR (Adverse Drug Reaction) [3]. In the United States, around 5% of all hospital admissions are the result of an ADR, and around 10%–20% of inpatients will have at least one ADR during their hospital stay [3], but the exact amount of deaths due to ADRs cannot be conformed because these patients have had other forms of severe health conditions.

ADRs can be basically divided into two categories as follows:

- Adverse Events – A negative experience encountered by an individual, this may or may not be due to the drug.

ADR is a special type of drug adverse event concern of the field known as pharmacovigilance. This refers to any injury occurring at the time a drug is used, whether or not it is identified as a cause of the injury.

- Serious Adverse Events – Any adverse event that is fatal, life threatening or results in hospitalization [21].

- Serious adverse drug reactions are classified which outcome is one of following, Death
- Life-threatening
- Hospitalization (initial or prolonged)
- Disability - significant, persistent, or permanent change, impairment, damage or disruption in the patient's body function/structure, physical activities or quality of life.
- Congenital abnormality
- Requires intervention to prevent permanent impairment or damage

Pharmacovigilance

Pharmacovigilance commonly known as drug safety, PV or PHV is the collection, detection, assessment, monitoring and prevention of adverse effects of pharmacological drugs [5]. The disaster caused by the drug thalidomide in 1961 [4] which made infants congenitally deformed due to the exposure of unsafe medicine taken by pregnant mothers caused the international community to make systems that address drug safety issues systematically.

The international communities took steps to make standardized adverse event reporting systems so that every incident occurring will be captured and made available under a centralized location.

Although a drug goes under heavy clinical trials that mimic the real-world drug usage, there are considerable differences between the two scenarios which brings the need for pharmacovigilance practices.

Although the physicians are supposed to report possible drug ADRs it is rare that a patient comes back and reports it to the physician, hence most of the drug ADRs go unreported which can be very dangerous. Finding these unreported adverse events is a major challenge for the pharmaceutical companies, but lately people tend to post these extensively on social media [6]. Thus, pharmaceutical companies must analyze through unstructured data like Facebook comments, tweets, call center records of health insurance companies and CRM (Customer Relationship Management) systems of pharmaceutical company sales representatives etc.

The data that is available in social media is highly important as it provides an interaction with the patient in a personal level, but this data is massive and manual evaluation to track down ADRs is a very labor-intensive task. Machine-learning based approaches can be used to analyze the data and flag the data that may potentially contain adverse events which can be further analyzed by professionals to confirm, this way it will reduce the human error that happens when going through that amount of data and it will also reduce the amount of expert labor required. The FDA states their regulation as an ADR should be reported within 24 hours of occurring [7], but with these manual methods achieving this timeline is impossible.

1.3 Motivation

Most of the work described in this thesis was conducted at IQVIA Solution Lanka (IQVIA) in Colombo, Sri Lanka. The reason for conducting the research is to solve a problem in their existing system called AETracker which is online system classifying drug adverse events manually.

So intention is to introduce a machine learning solution to classify Drug Adverse Event and Non-Drug Adverse Events.

Based on social web data, the system automatically extracts drug induces and medical terms for Drug Adverse Events and Non-Drug Adverse Events, and the match between these features. It then uses various machine learning methods to predict the success of classifying adverse events. The overarching research question for this paper is to Classifying Drug Adverse Events using Social Media Data. By proposing the first system to ADE at an early stage, the main contributions of this research are in two areas: In the first place, this work exhibits how uninhibitedly accessible information of various kinds (counting organized information and unstructured information) can be gathered, intertwined, and broke down to prepare AI calculations. When planning and creating data framework curios, such information-based methodologies can give amazing figures and suggestions to help business choices. As far as we could possibly know, we are additionally the first to use such information and banner the tweets. Second, our exploration proposes a few novel highlights, for example, dynamic system highlights, plot theme dispersions, the match among ADE and non-ADE at beginning periods. We indicated that these highlights all make extraordinary commitments to the framework's presentation and help to clarify significant factors in Drug Adverse Events.

1.4 Objectives

As a solution companies like IQVIA [8] has come up with products like AE Tracker which scans the social media and forums and blogs, but they use a medical ontology to detect adverse events from social media, this approach has a very high false positive rate as well as the ontology should be kept up to date.

There is a considerable research effort in classifying whether a social media content is an adverse event or not but the automatic extraction of this adverse effect from the social media content is not much done.

In this research my target is to resolve following problems. So as to chronicle the primary goal, there are some other objectives,

O1. Extract ADRs (Adverse Drug Reactions) with accuracy rates from large continuous streams of social media data.
O2. Use continuously update the model to capture ADR's from Social media accurately.
O3. Implement a deep learning solution model that gathers data from all types of social media and other types of data such as forums blogs and automatically detect possible drug adverse reactions from these.

1.5 Scope

S1. Analysis will be done for extract ADRs with low false positive rates from large continuous streams of social media data
S2. Analysis will be done for learning to continuously update the model to capture ADR's from Social media accurately
S3. Implement a modal to gather social media and forum data to detect possible drug adverse reactions
S4. Capture the drug adverse pattern
S5. Plot reports of the drug adverse events from their contents

Chapter Two

Related Research or Previous Work

With increasing popularity of social media, health care blogs and forums where people post about their health care issues and medical reports that are electronically available researching on ADR detection from digital data has gained attention. The recent developments of advanced machine learning is a reason for research interests in this field.

The following sections summarize different aspects of using Machine on detection of ADR's from Social Media and details about past research on these areas.

The essential components for this process are a data collection layer which gathers data from social media & other types of data, a data pre-processing stage should remove any unwanted contents such as advertisements, URL's & links in contents and images. The cleaned data should be sent to a trained model where it first classifies the data into ADR & non-ADR content first and then extraction of the ADR contents should take place. [4,5]

Social media data contains a lot of slang, symbols emojis etc. These need to be filtered out before extracting features to determine ADR. Apart from this there are general preprocessing techniques that can be used when text mining. Stop words are usually removed hence they do not provide meaning when taken individually, but past research on ADR detection from Social media data suggests that when taking word 2-grams and 3-grams as features stop words has a positive effect [9].

POS (Part of Speech) tagging is also done as a pre-processing step, here POS tags are assigned to each word and these tags are then used for classification, since the language used in twitter is different from the usual written language, POS tagging specific to online conversations [10] has been used in past research. We can also use general POS taggers such as Stanford POS tagger [11], it is a lexicalized probabilistic parser which provides various information such as the syntactic structure of text segments, dependencies and POS tags.

Other preprocessing techniques involve tagging text using external vocabularies such as UMLS (Unified Medical Language System) [12], removing social media specific entities such as hashtags and special characters such as @ symbols, removing possible advertisements since data from Social media contains a lot of advertisements and marketing material and stemming and lemmatization to get the base forms of words [13].

Social media generates data at a very fast rate, identifying whether a piece of text contains an ADR or not is the first step in mining ADRs from social media, for this popular approach has been to use supervised classifiers that use different features related to the task in hand, a research done by Abeed et al. [9] uses Naïve Bayes, SVM (Support Vector Machines) and ME (Maximum Entropy) classifiers to classify tweets into ADR and non ADR. SVM has been the natural choice for text classification because of its ability to deal with high dimensional feature space.[14] has used a SVM and a MNB (Multinomial Naïve Bayes) classifier for binary classification of comments from DailyStrength which is a healthcare related forum, [12] has used SVM, MNB, ME and a tree based classifier known as J48 to classify tweets containing possible mentions of drug abuse and then used stacking, a technique where the predictions from the different classifiers are combined and another algorithm is trained to make a final decision based on the individual predictions. Most of the past research for binary text classification has either been using SVM or the ME gclassifier or an ensemble of both using a weighted average [13,15]. UMLS is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. A key problem when mining ADRs from text is that distinguishing mentions about adverse reactions (an unexpected, negative effect resulting from the adequate use of the drug) from mentions about indications (the sign or symptom for which the drug was prescribed in the first place). This research has considered this factor when doing the annotation.

In 2017,Kathy Lee and Ashequl Qadir developed CNN(Convolutional Neural Network) model for ADE classification with various unlabeled data sets in tweets, and evaluated a model on the Twitter data set used in the PSB 2016 Social Media Shared Task .Without large no of dataset it can performed at some extend[22].In 2017,Ahmad P Tafti and Jonathan Badger did investigate a big data machine learning strategy for ADE discovery on massive dataset from PubMed Central and social media. This contribution illustrates possible capacities in big data biomedical text analysis using advanced computational methods with real-time update from new data published on a daily basis [23].

Chapter Three

Problem Analysis and Methodology

3.1 Problem Analysis

Breaking down the research problem is a major part of this research. 4593 tweets which extracted in November 2019 included for this analysis. All the tweets are cleaned by removing twitter Return handles (RT @xxx:), twitter handles (@xxx), URL links (httpxxx), URL links (httpxxx) and special characters, numbers, punctuations (except for #).

1627 tweets removed as duplicate values.2846 tweets remained with unique but with different languages like French, German, Chinese, Philippines etc. All of them translated to English language (See Appendix A). using google translator.

After clean, preprocessed and transformed tweet data set summary is below with 2846 tweets in Table 1:

Column (Attribute)	Description
Drug	Drug name
User	Creator of the tweet
Tweet	Main text in the tweet
Flag	Classified label
No of followers	No of followers for the tweet

Table 1 Column Summary

Please find the drug and medical application list which used to create this dataset in Figure 1,

Anesthesia	Antihistamine	Antipsychotic	Aspirin	Atenolol	Atorvastatin
Azithromycin	Dettol	Dexamethasone	Diazepam	Dopamine	Durex
Ephedrine	Gabapentin	Galantamine	Heparin	Ibuprofen	HPV
Lamotrigine	Lorazepam	Melatonin	Meloxicam	Merck	Metformin
Methylphenidate	MSD	mucinex	nurofen	ondansetron	Orlistat

Sildenafil	Statins	Veet	Vioxx	Warfarin	Wellbutrin

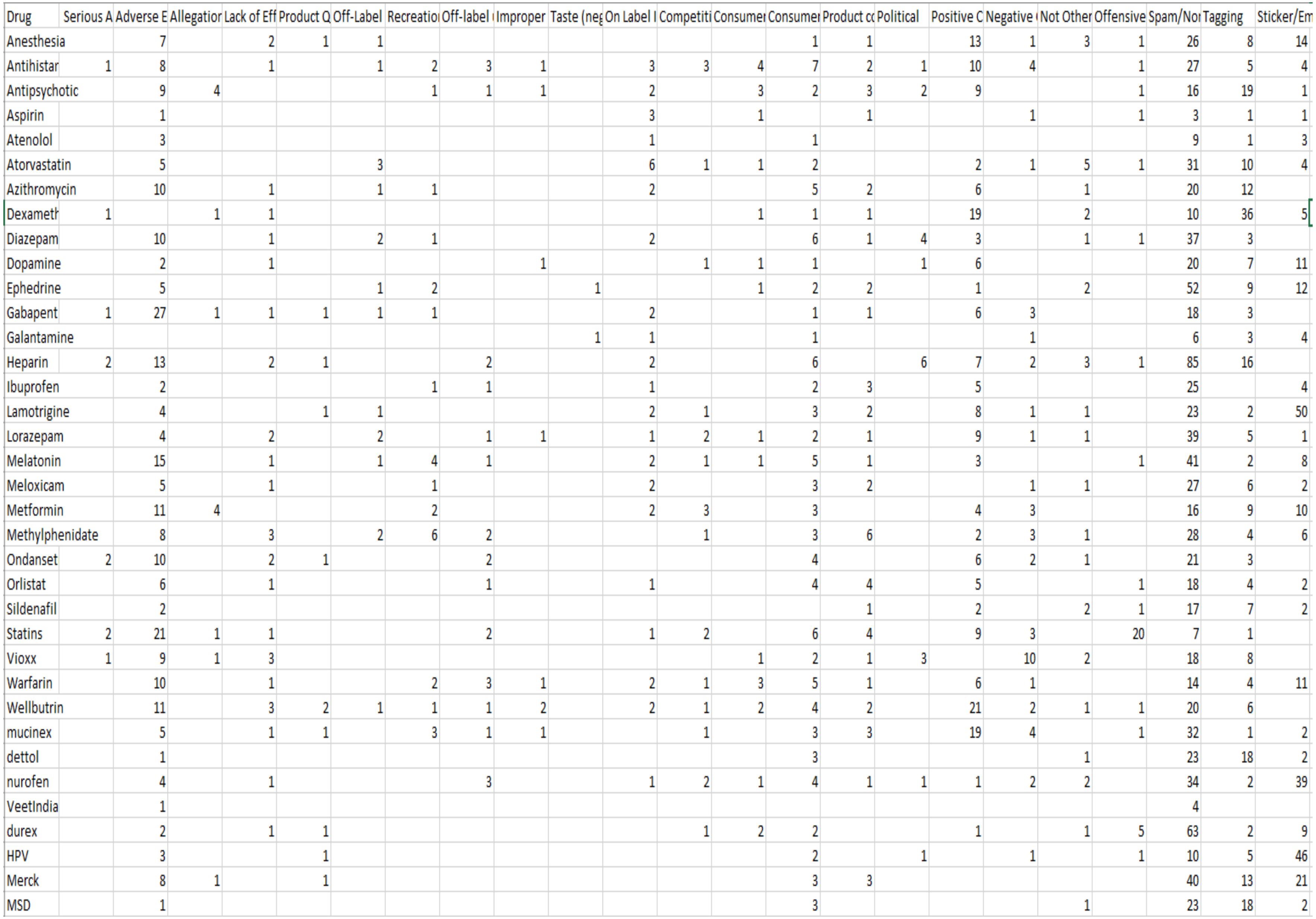
Figure 1 Drug List

All these tweets are manually verified by for generic 23 classification flags used in pharmacology industry. Such as shown in the Table 2,

RSI (Reporting to the client)	*No RSI (Not reporting to the client)*
Serious Adverse Event	Spam/Nonsense
Adverse Event	Tagging
Allegation of Death	Offensive or abusive language
Lack of Efficacy	Sticker/Emoji
Product Quality Complaint	
Off-Label Usage	
Recreational drug use or abuse	
Off-label usage that requires HCP advice	
Improper Attribution of Product Benefits	
Taste (negative)	
On Label Incomplete	
Competitive Comparative Threat	
Consumer Questions/Suggestion (Not Channel-Specific)	
Consumer Question/Suggestion (Channel-Specific)	
Product Cost	
Political	
Positive Comment	
Negative Comment	
Not otherwise classified	

Table 2 Flag List

ICD 10 Medical drug induces terms [21] used for classifying these preprocessed texts. After manually verified the dataset summary is in Figure2.



Drug	Serious A	Adverse E	Allegatior	Lack of Eff	Product Q	Off-Label	Recreatio	Off-label (	Improper	Taste (neg	On Label I	Competiti	Consumer	Consumer	Product cc	Political	Positive C	Negative (	Not Other	Offensive	Spam/Nor	Tagging	Sticker/Em
Anesthesia		7		2	1	1								1	1		13	1	3	1	26	8	14
Antihistar	1	8		1		1	2	3	1		3	3	4	7	2	1	10	4		1	27	5	4
Antipsychotic		9	4				1	1	1		2		3	2	3	2	9			1	16	19	1
Aspirin		1									3		1		1			1		1	3	1	1
Atenolol		3									1			1							9	1	3
Atorvastatin		5				3					6	1	1	2			2	1	5	1	31	10	4
Azithromycin		10		1		1	1				2			5	2		6		1		20	12	
Dexameth	1		1	1									1	1	1		19		2		10	36	5
Diazepam		10		1		2	1				2			6	1	4	3		1	1	37	3	
Dopamine		2		1					1			1	1	1		1	6				20	7	11
Ephedrine		5				1	2			1			1	2	2		1		2		52	9	12
Gabapent	1	27	1	1	1	1	1				2			1	1		6	3			18	3	
Galantamine										1	1			1				1			6	3	4
Heparin	2	13		2	1			2			2			6		6	7	2	3	1	85	16	
Ibuprofen		2					1	1			1			2	3		5				25		4
Lamotrigine		4			1	1					2	1		3	2		8	1	1		23	2	50
Lorazepam		4		2		2		1	1		1	2	1	2	1		9	1	1		39	5	1
Melatonin		15		1		1	4	1			2	1	1	5	1		3			1	41	2	8
Meloxicam		5		1			1				2			3	2			1	1		27	6	2
Metformin		11	4				2				2	3		3			4	3			16	9	10
Methylphenidate		8		3		2	6	2				1		3	6		2	3	1		28	4	6
Ondanset	2	10		2	1			2						4			6	2	1		21	3	
Orlistat		6		1				1			1			4	4		5			1	18	4	2
Sildenafil		2													1		2		2	1	17	7	2
Statins	2	21	1	1				2			1	2		6	4		9	3		20	7	1	
Vioxx	1	9	1	3									1	2	1	3		10	2		18	8	
Warfarin		10		1			2	3	1		2	1	3	5	1		6	1			14	4	11
Wellbutrin		11		3	2	1	1	1	2		2	1	2	4	2		21	2	1	1	20	6	
mucinex		5		1	1		3	1	1			1		3	3		19	4		1	32	1	2
dettol		1												3					1		23	18	2
nurofen		4		1				3			1	2	1	4	1	1	1	2	2		34	2	39
VeetIndia		1																			4		
durex		2		1	1							1	2	2			1		1	5	63	2	9
HPV		3			1									2		1		1		1	10	5	46
Merck		8	1		1									3	3						40	13	21
MSD		1												3					1		23	18	2

Figure 2 Flagged Data Set

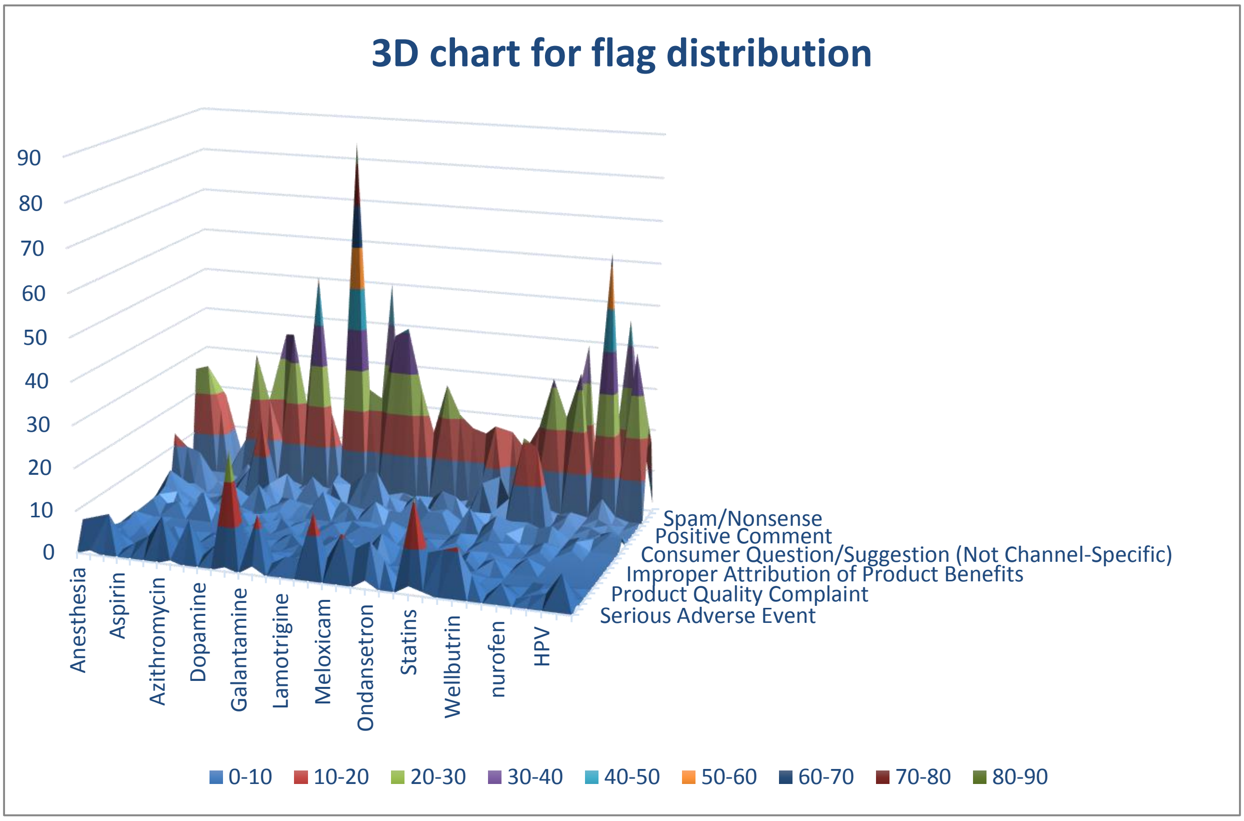


Figure 3 3D Data Distribution

For each drug/medical appliance is mapped to a specific flag is described in Figure 3.

missing values for the flagged data set shown as below in Figure 4, According to this chart lot of missing values are from Taste (Negativity) and adverse event flag has few missing values. As we want to analyzing mostly for the adverse events this data set is enriched with most adverse and serious adverse event flags. Most available flag is Spam/Nonsense.

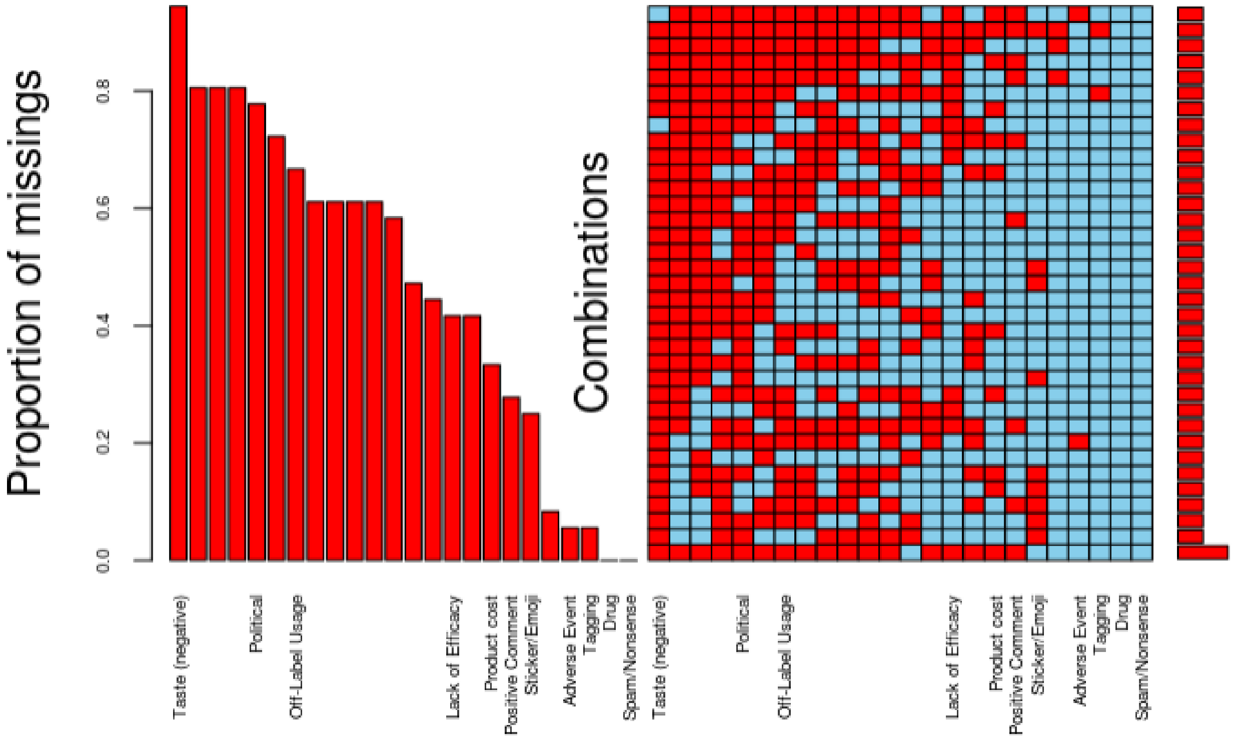
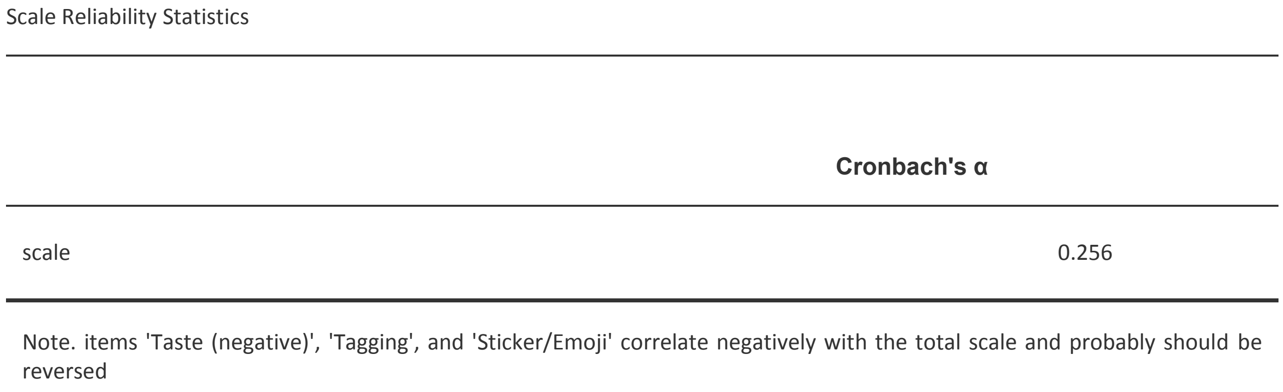


Figure 4 Missing Values

In our dataset there are 23 variables. So reliability of each and every variable is calculated. Reliable analysis done for this dataset as scale analysis, we are using cronbach's alpha [16-19] as scale statistics and found following result. Cronbach's Alpha is the most commonly used statistic for determining the internal consistency of measurements. Reliability statics analysis in Table 3.



Scale Reliability Statistics

	Cronbach's α
scale	0.256

Note. items 'Taste (negative)', 'Tagging', and 'Sticker/Emoji' correlate negatively with the total scale and probably should be reversed

Table 3 Cronbach's Result

After reversed scale with Taste(negative), Tagging and Sticker/Emoji flags relatability analysis result in table 4 ,

Scale Reliability Statistics

	Cronbach's α
scale	0.360

Table 4 Cronbach's reverse scale result

In this research we are doing a principal component analysis for reduce the dimension in the feature selection. varimax rotation was used in Table 5.

Component Loadings

	Component	
	1	Uniqueness
Serious Drug Adverse Event	0.501	0.749
Adverse Event	0.654	0.572
Allegation of Death		0.986
Lack of Efficacy	0.662	0.562
Product Quality Complaint		0.943
Off-Label Usage		0.927
Recreational drug use or abuse	0.426	0.818
Off-label usagthatrequirs HCPadvc	0.732	0.465
Improper Attribution of Prdct Bnfts	0.497	0.753
Taste (negative)		0.914
On Label Incomplete		0.919
Competitive Comparative Threat	0.568	0.678

CnsmrQstn /Sggstn (NtChnnl-Spcfc)	0.445	0.802
ConsumrQstn / Sggstn(Chnnl-Spcfc)	0.718	0.484
Product cost	0.405	0.836
Political	0.300	0.910
Positive Comment	0.533	0.716
Negative Comment	0.518	0.732
Not Otherwise Classified		1.000
Not Otherwise Classified	0.347	0.879
Spam/Nonsense		0.928
Tagging		0.985
Sticker/Emoji		0.967

Note. 'varimax' rotation was used

Table 5 Principal Component Analysis

As the high dimensionality of the dataset, PCA (Principal Component Analysis) is used to reduce the dimensionality, which act as summaries of features. We are calculating each eigenvalue as descending order and taking 90% of variance in eigenvalues in Table 6.

Component	Eigenvalue	% of Variance	Cumulative %
1	4.4746	19.455	19.5
2	2.5084	10.906	30.4
3	2.1720	9.444	39.8
4	1.9237	8.364	48.2
5	1.7634	7.667	55.8
6	1.5020	6.531	62.4

7	1.3316	5.790	68.2
8	1.1613	5.049	73.2
9	1.1229	4.882	78.1
10	0.9764	4.245	82.3
11	0.9384	4.080	86.4
12	0.6078	2.642	89.1
13	0.5601	2.435	91.5
14	0.4404	1.915	93.4
15	0.3477	1.512	94.9
16	0.2931	1.274	96.2
17	0.2375	1.032	97.2
18	0.1960	0.852	98.1
19	0.1378	0.599	98.7
20	0.1300	0.565	99.2
21	0.1024	0.445	99.7
22	0.0427	0.186	99.9
23	0.0298	0.129	100.0

Table 6 Initial Eigenvalues

Figure 5 describes how each eigen value is change with different variable flags.

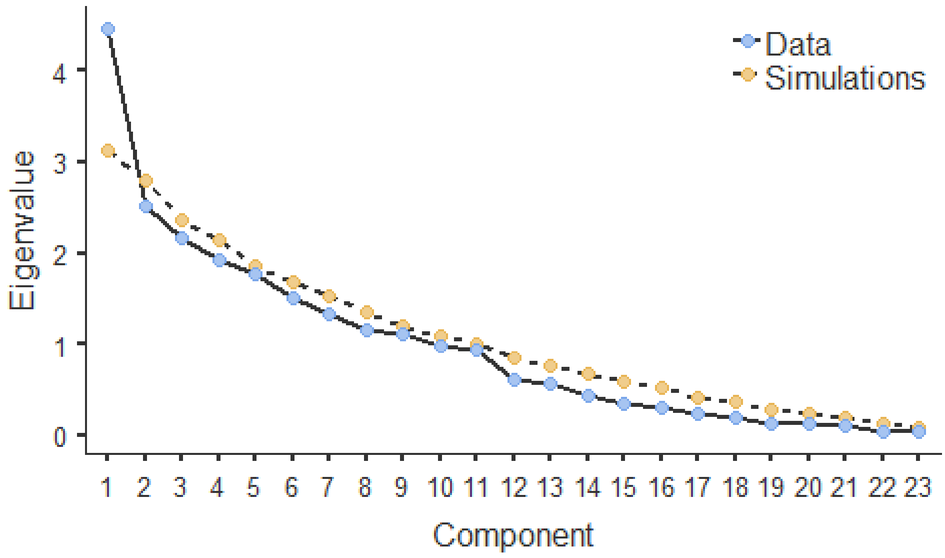


Figure 5 Screen Plot

According to the above result, we are taking following flags as principal components such as Serious drug adverse events, Adverse event , Allegation of death , Lack of efficacy , Product quality complaint , Off-Label usage , Recreational drug use or abuse , Off-label usage (HCP advice needed) , Improper attribution of product benefits , Taste (negative) , On label incomplete , competitive comparative threat and Consumer question/suggestion (Not channel specific).

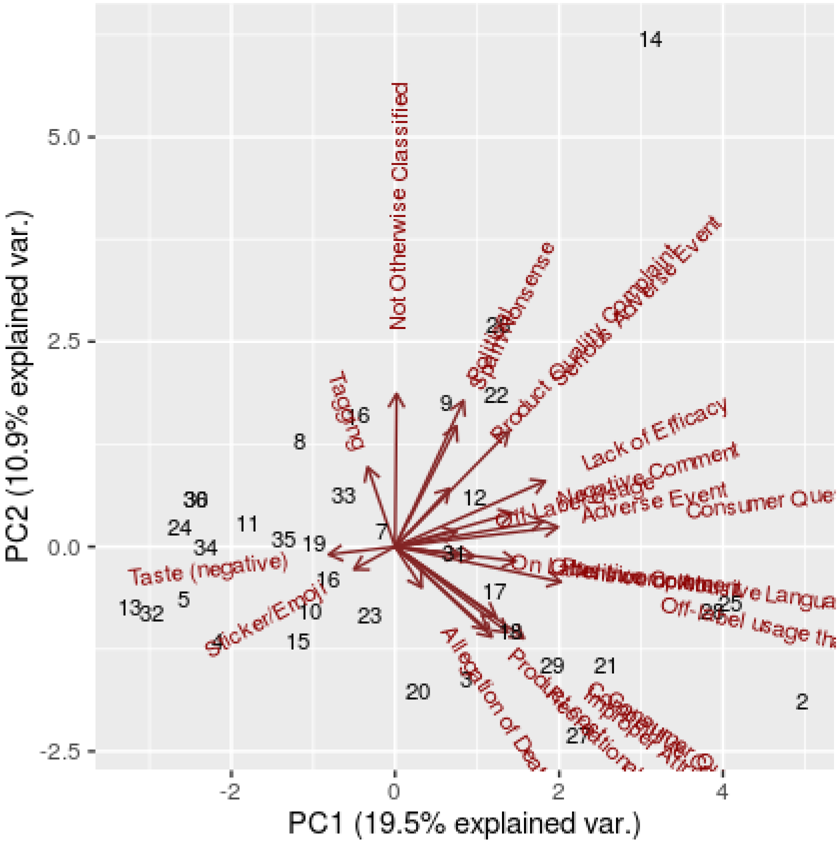


Figure 6 PCA Plot

Figure 6 describes every variables scattered through the multi-dimensional space.

3.2 Methodology

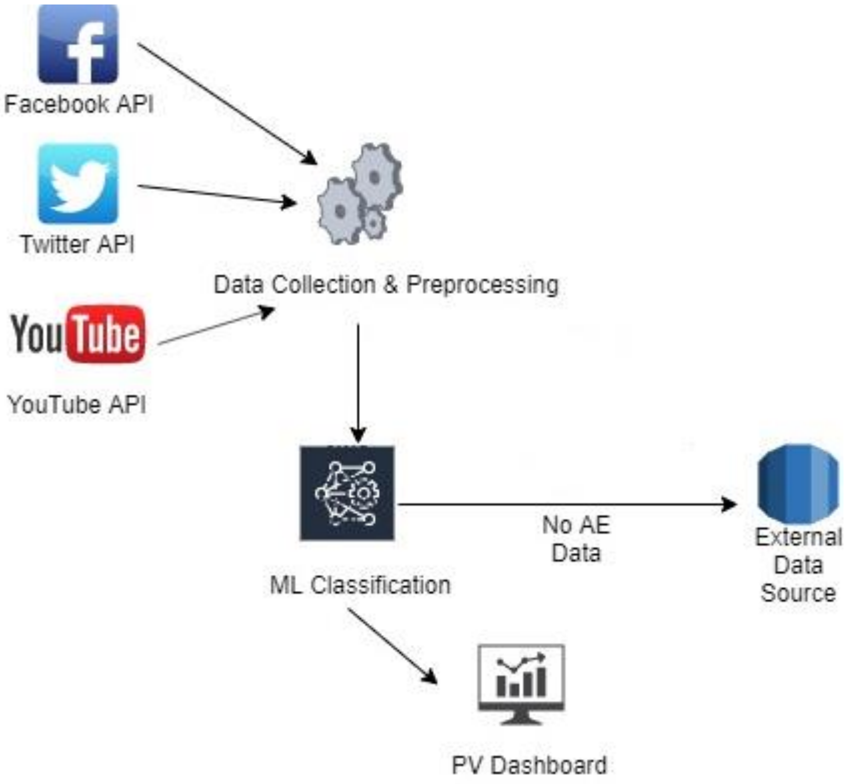


Figure 7 Proposed System Architecture

Shown in figure 7 is the proposed high-level system architecture. The data collection and preprocessing layer will interact with the different Social Media API's such as Facebook, Twitter, pull in data of interest to the AE tracker system, the API's give us the option to follow keywords of interest and get relevant Social Media postings containing these keywords. It then preprocesses the data removing images, advertisements etc. The data collection layer also formats the messages received into a common format, ideally JSON containing the content, a URL of the posting and the source it came from. This formatted message will be sent on to the data classification layer.

The Data classification layer will be a machine learning based text classification process. This layer will classify the streams of texts into ADR's and Non-ADR's and output it to an external data store as well as to the drug ADR relationship extraction layer.

The Drug ADR relationship extraction layer will extract drugs and adverse events relationships from Drug ADR postings. This will make use of a neural network and phrase embeddings as features, we decided to use phrase embeddings as our literature review showed us that phrase embeddings show relationships between different entities such as drug mentions, disease names and ADR's.

Chapter Four

Results

K - means clustering

As our research main method is to be clustering the drug adverse events. We used K-means unsupervised learning methods having an iterative process in which the dataset are grouped into k number of predefined non-overlapping clusters or subgroups making the inner points of the cluster as similar as possible while trying to keep the clusters at distinct space it allocates the data points to a cluster so that the sum of the squared distance between the clusters centroid and the data point is at a minimum, at this position the centroid of the cluster is the arithmetic mean of the data points that are in the clusters[20].

Used Elbow method and plotted the below graph.

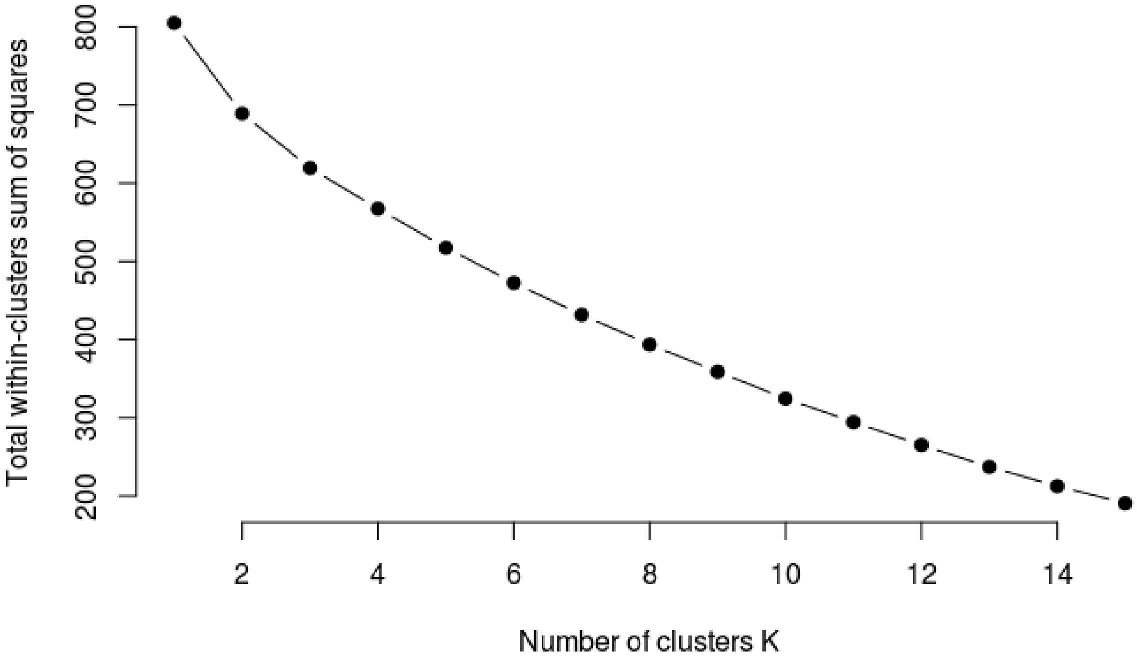


Figure 8 Elbow Graph

Used Elbow method and plot the below graph. Found 3 is the optimum cluster values K. According to the information gain plotted the below graph.

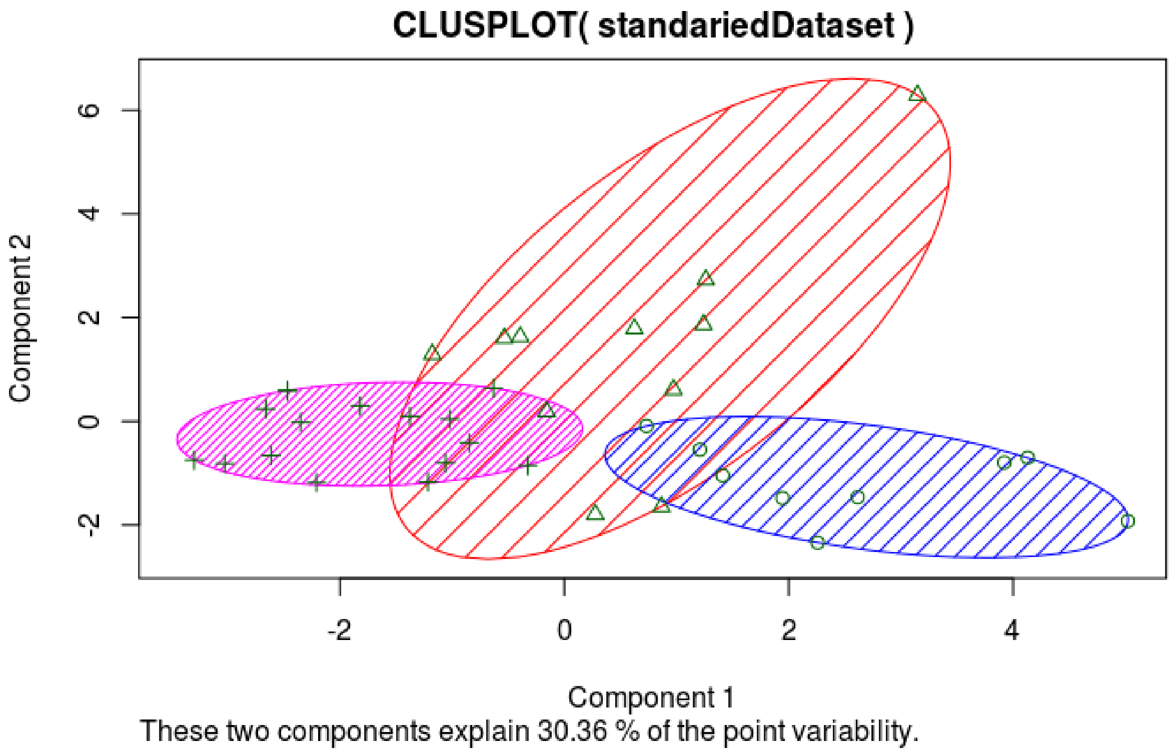
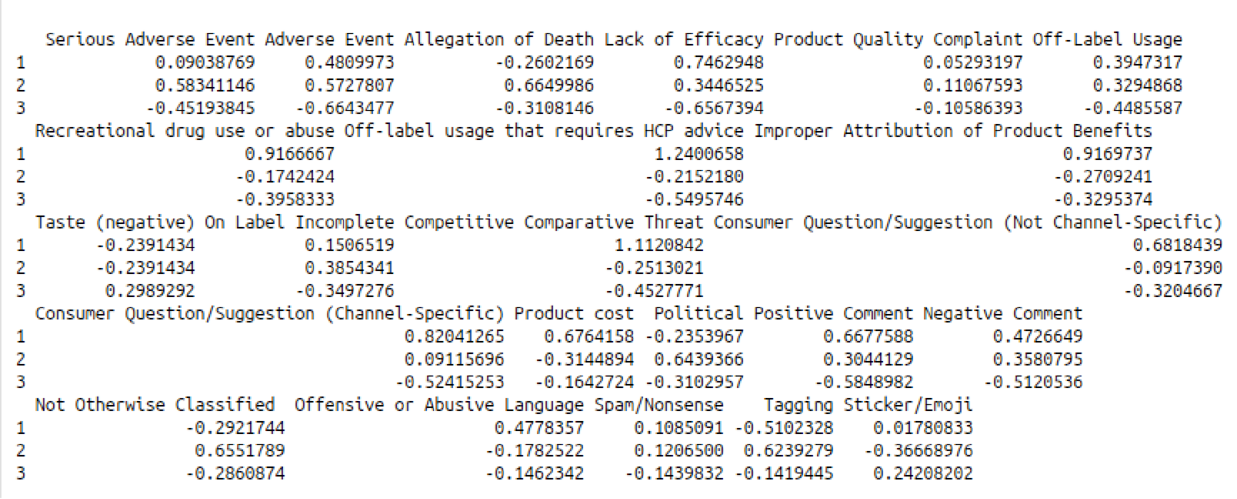


Figure 9 Cluster Graph

Three clusters have identified in Figure 9.



```
  Serious Adverse Event Adverse Event Allegation of Death Lack of Efficacy Product Quality Complaint Off-Label Usage
1            0.09038769     0.4809973          -0.2602169        0.7462948                0.05293197       0.3947317
2            0.58341146     0.5727807           0.6649986        0.3446525                0.11067593       0.3294868
3           -0.45193845    -0.6643477          -0.3108146       -0.6567394               -0.10586393      -0.4485587
  Recreational drug use or abuse Off-label usage that requires HCP advice Improper Attribution of Product Benefits
1                      0.9166667                                1.2400658                                0.9169737
2                     -0.1742424                               -0.2152180                               -0.2709241
3                     -0.3958333                               -0.5495746                               -0.3295374
  Taste (negative) On Label Incomplete Competitive Comparative Threat Consumer Question/Suggestion (Not Channel-Specific)
1       -0.2391434           0.1506519                      1.1120842                                           0.6818439
2       -0.2391434           0.3854341                     -0.2513021                                          -0.0917390
3        0.2989292          -0.3497276                     -0.4527771                                          -0.3204667
  Consumer Question/Suggestion (Channel-Specific) Product cost  Political Positive Comment Negative Comment
1                                      0.82041265    0.6764158 -0.2353967        0.6677588        0.4726649
2                                      0.09115696   -0.3144894  0.6439366        0.3044129        0.3580795
3                                     -0.52415253   -0.1642724 -0.3102957       -0.5848982       -0.5120536
  Not Otherwise Classified  Offensive or Abusive Language Spam/Nonsense    Tagging Sticker/Emoji
1               -0.2921744                      0.4778357     0.1085091 -0.5102328    0.01780833
2                0.6551789                     -0.1782522     0.1206500  0.6239279   -0.36668976
3               -0.2860874                     -0.1462342    -0.1439832 -0.1419445    0.24208202
```

Figure 10 Cluster Summary

Figure 10 described how each centroid towards for each variable.

According to the Figure 9, 3 clusters identified. Such as,

cluster 1 : Adverse events, allegation of death, recreational drug use or abuse, off-label usage that requires HCP advice, Improper attribution of product benefits, On Label Incomplete, competitive comparative threat, not channel-specific, product cost, positive comment, off-label usage cluster 2 : Serious Adverse events, lack of efficacy, product quality complaint, channel-specific, political, negative comment, not otherwise classified, offensive or abusive language, spam/nonsense

cluster 3 : taste negative, tagging, sticker/emoji

we named this clusters as,
Cluster 1 -> ADR (Adverse Drug Events)
Cluster 2 -> Toward ADR (Toward Adverse Drug Events)
Cluster 3 -> No ADR (No Adverse Drug Events)

SOM (Self Organizing Map)

For dimensional reduction we used self-organizing map.

There are 23 variables in my dataset. Draw the self-organizing map according to the result from principal component analysis. So, we are using 14 principal components.

This is the training chart with default alpha values in Figure 11.

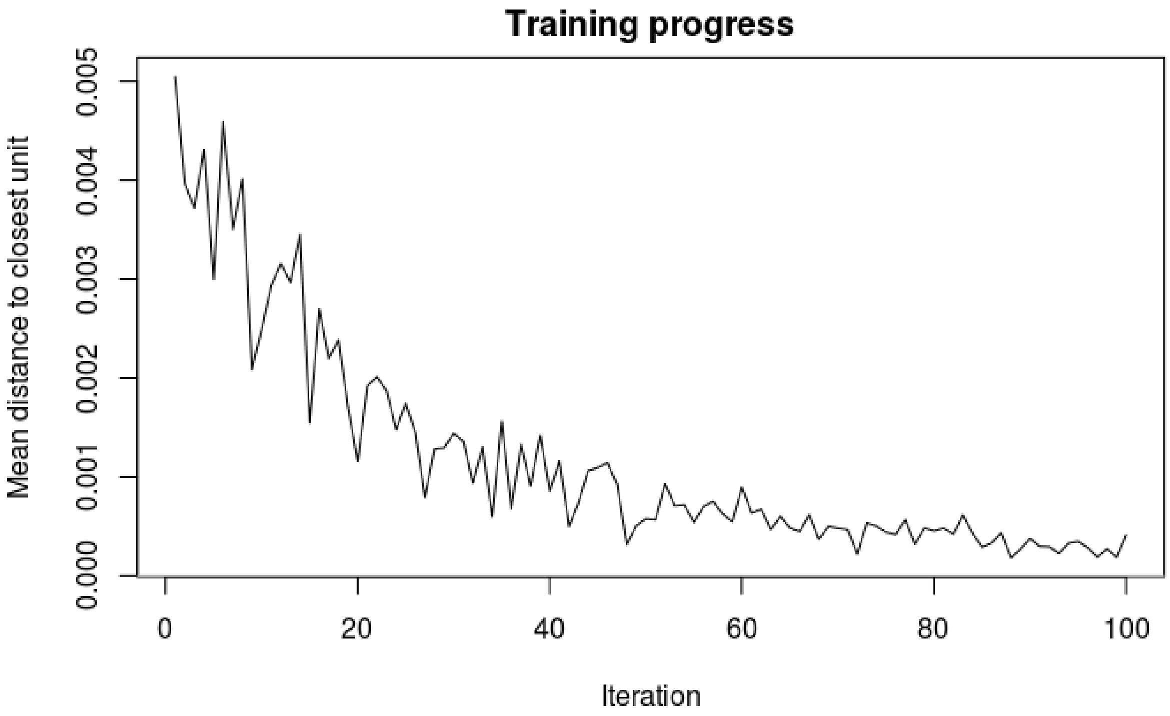


Figure 11 Training Progress

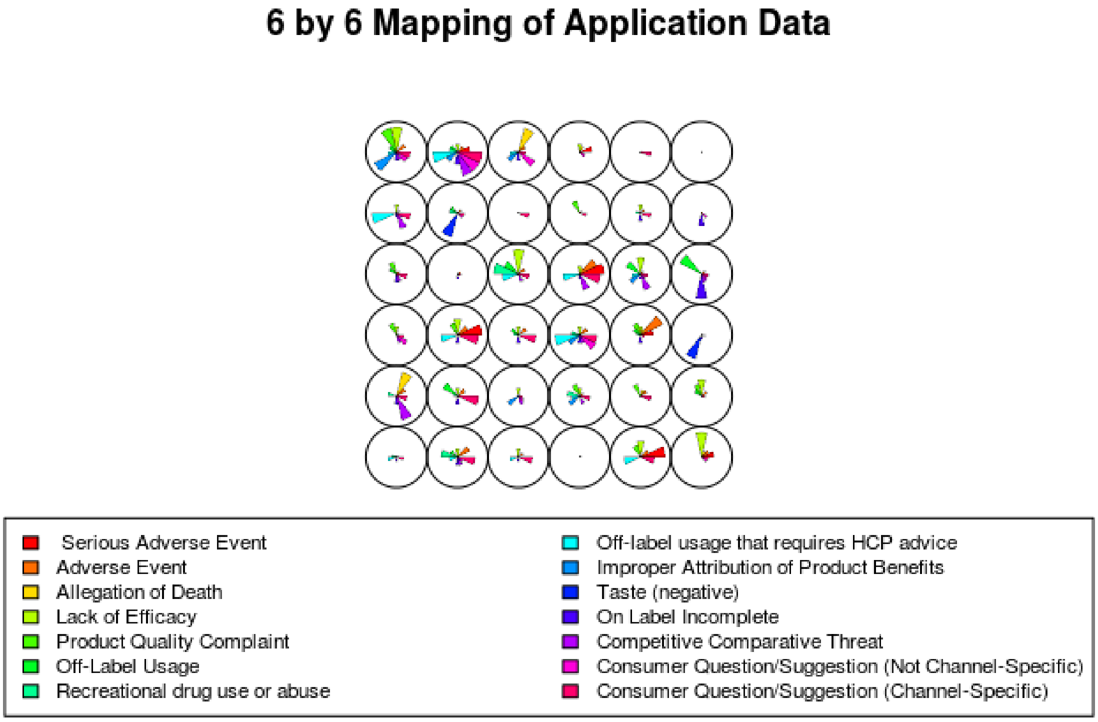


Figure 12 Application Data

The default visualization of the weight vectors is a “fan diagram”, where individual fan representations of the magnitude of each variable in the weight vector is shown for each node in Figure 12.In here we can see each principal components distribution in the dataset.

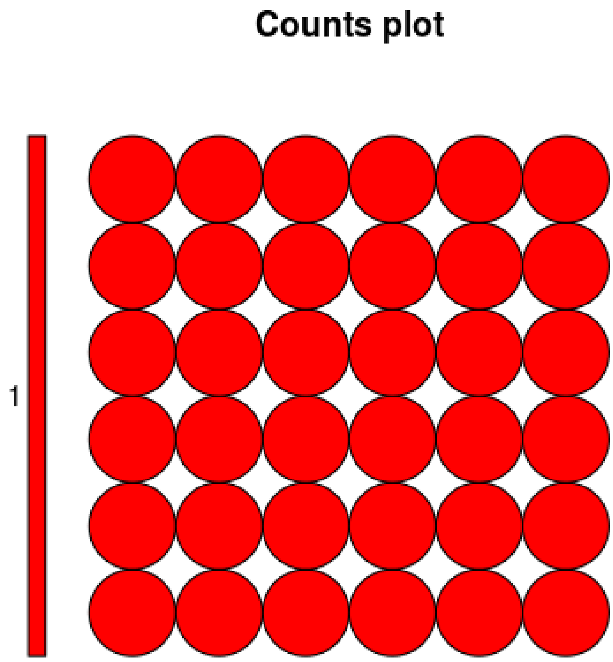


Figure 13 Counts Plot

Count plot in Figure 13,we see every drug in evenly spread in all variables.

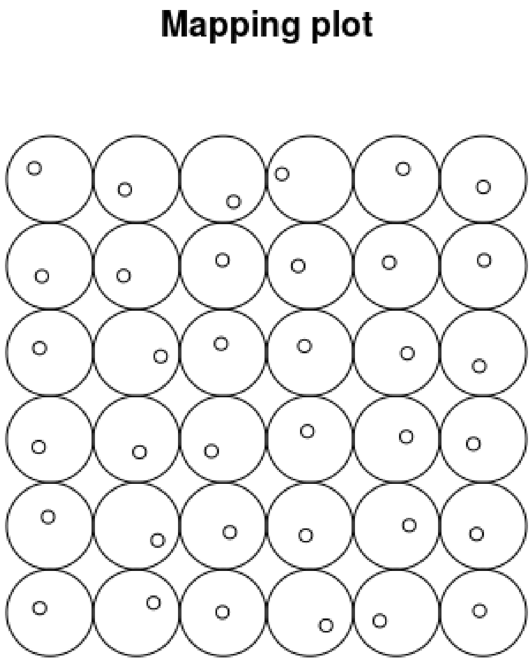


Figure 14 Mapping Plot

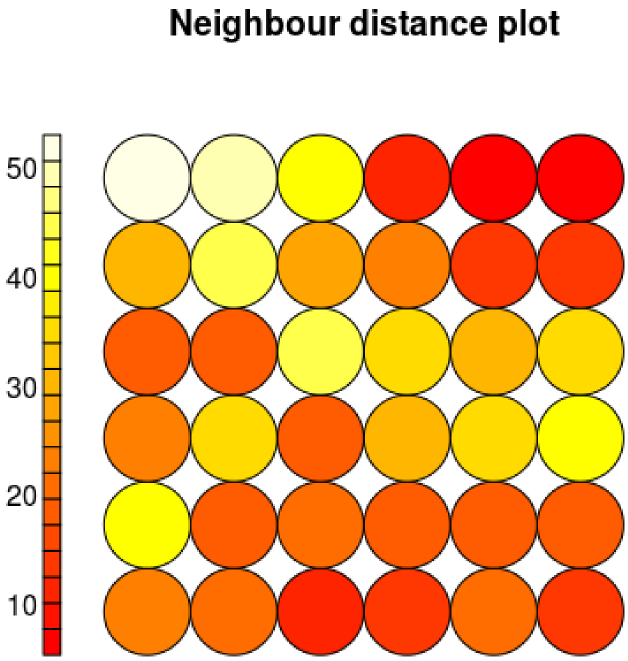


Figure 15 Neighbor Distance Plot

Overall distance to each variables in the SOM(Self organized map) is showed in Figure 14-15.According to this plot, most of variables are closed to each other.

Chapter Five

Validation and Evaluation

To evaluate the performance of our un-supervised clustering model, the trained model applied to each tweet in the 20% held-out test data to predict a binary label (Agree or Not Agree).

Table 7 Example of False Negatives, False Positives and Prediction Disagreement

Tweet	Label	prediciton model	Agree/Not Agree
Nope its not. More deaths in the placebo arm. That's the reason for the use of Statins. Thanks for confirming the same.	Serious Adverse Event	Toward ADR	Not Agree
Merck has killed hundreds of thousands of Americans with Vioxx and other dangerous drugs. They have paid billions of dâ€¦	Serious Adverse Event	Toward ADR	Not Agree
Did you know that even after regular cleaning, the surfaces in your home could have illness causing germs? On 16th December, join the movement against germs, because #CleanIsNotGermFree. Know more: #Dettol	Adverse Event	ADR	Agree
Another fake profile at its finest ... Insanity ... I have found another one for you for the next contraceptive advertisement ... So that people no longer arise ...	Adverse Event	ADR	Agree
Monotherapy with Metformin versus Sulfonylureas and Risk of Cancer in Type 2 Diabetic Patients: A Systematic Review and Meta-Analysis.	Allegation of Death	ADR	Agree
30 years as a nurse tell me he only has 5 years left. Death is a great equalizer especially if take your statins with KFC.	Allegation of Death	ADR	Agree
The time has come. In under an hour Iâ€™m getting my wisdom teeth out. First thing Iâ€™m doing when I sit down for this surgery is asking my surgeon if I can blast new uzi before going under anesthesia. #gasssssssss	Lack of Efficacy	Toward ADR	Agree
iâ€™m just going to take an antihistamine... why canâ€™t i just have normal skin	Lack of Efficacy	Toward ADR	Agree
I hate hate hate hate being sick. This climate change is going to kill me. I got sick twice in 9 days. Going on day 10. Theraflu, DayQuil, mucinex, Dominican cough syrup, vaporub, lemon tea, chamomile, lemon zinger, lemon water and honey. Lord help me ðŸ˜	Product Quality Complaint	Toward ADR	Agree
Need something stronger pal. IV ondansetron Stat	Product Quality Complaint	Toward ADR	Agree
I stopped listening to my doctor when he insisted I take Atorvastatin.'cos I can read innit!	Off-Label Usage	ADR	Agree
Next time try 3 days of Azithromycin works like a bomb	Off-Label Usage	ADR	Agree
Instead of kids selling Lucozade bottles and and Kit Katâ€™s they gonna be selling Paracetamol and Ibuprofen	Recreational drug use or abuse	ADR	Agree
ROLLER SKATING LEADS TO HARD DRUGS LIKE COCAINE AND MELATONIN	Recreational drug use or abuse	ADR	Agree

The combination I currently use is guaifenesin (Mucinex) for congestion, ibuprofen (Advil) for fever + pain, and cough drops for sore throat. If you're coughing a lot, you can use numbing drops like CÄ“pacol along with tea with honey (buckwheat honey has been shown to help).	Off-label usage that requires HCP advice	ADR	Not Agree
I saw three Nurofen posts where the influencers had alcohol. Maybe im not clued up but theoretically, can you drink painkillers with alcohol?ðŸ¤¨	Off-label usage that requires HCP advice	ADR	Not Agree

Table 7 presents some example tweets for which our model had prediction errors or there were prediction disagreements. Based on them 12.5% is false positive and 12.5% is false negative.

False Negative:

First and second tweets are classified as Serious Drug Adverse Events because of the word death and our model predict them as Toward ADR. But getting the sense of meaning of that phrase they can be classified as political, spam/nonsense or negative comments.

False Positive:

Last two tweets in Table 1 classified as Off-label usage that requires HCP advice because of certain drug names and some medical terms use there. But first tweet from there should be consumer question/suggestion (channel – specific) and other one is recreational drug use of abuse.

Chapter Six

Conclusion and Future Work

Apparently, this work is the first to research a big data machine learning strategy for ADE discovery on massive dataset downloaded from social media website. This commitment delineates potential limits in text analysis using machine learning methods with real-time update from new data published on a daily basis.

References

[1] En.wikipedia.org. 2020. Medicine. [online] Available at:

<https://en.wikipedia.org/wiki/Medicine> [Accessed 14 June 2020].

[1] Talbot, J C, and B S Nilsson. “Pharmacovigilance in the pharmaceutical industry.” *British journal of clinical pharmacology* vol. 45,5 (1998): 427-31. doi:10.1046/j.1365-2125.1998.00713.x

[2] Golder S.A., Wilkinson D.M., Huberman B.A. (2007) Rhythms of Social Interaction: Messaging Within a Massive Online Network. In: Steinfield C., Pentland B.T., Ackerman M., Contractor N. (eds) Communities and Technologies 2007. Springer, London

[3] Stephanie N. Schatz, Pharm.D, BCPS and Robert J. Weber, Pharm.D, BCPS “Adverse Drug Reactions”

[4] “Thalidomide,” Science Museum. [Online]. Available: http://broughttolife.sciencemuseum.org.uk/broughttolife/themes/controversies/thalidomide. [Accessed: 16-May-2020].

[5] “Pharmacovigilance,” World Health Organization, 20-Nov-2015. [Online]. Available: https://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/. [Accessed: 16-May-2020].

[6] “PharmaTech.com” Wego Health http://www.pharmexec.com/pharma-companies-can-solve-social-media-adverse-events-reporting-problem-and-stop-worrying /. [Accessed: 16-May-2020].

[7] “FDA Drug Topics: An Overview of Pharmacovigilance in the Center for Drug Evaluation and Research (CDER)” [Online]. Available: https://www.fda.gov/media/122835/download. [Accessed: 16-May-2020].

[8] “Welcome to IQVIA - A New Path to Your Success Via Human Data Science,” IQVIA. [Online]. Available: https://www.iqvia.com/. [Accessed: 16-May-2020].

[9] Abeed Sarker & Graciela Gonzalez, “Portable automatic text classification for adverse drug reaction detection via multi-corpus training”, Department of Biomedical Informatics, Arizona State University

[10] Owoputi O, O’Connor B, Dyer C, Gimpel K, Schneider N, Smith NA. Improved part-of-speech tagging for online conversational text with word clusters.

[11] “The Stanford NLP Group,” The Stanford Natural Language Processing Group. [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml. [Accessed: 16-May-2020].

[12] “Unified Medical Language System (UMLS),” U.S. National Library of Medicine. [Online]. Available: https://www.nlm.nih.gov/research/umls/index.html. [Accessed: 16-May-2020].

[13] Bahadorreza Ofogh, Samin Siddiqu, and Karin Verspoor read-biomed-ss: “Adverse drug reaction classification of microblogs using emotional and conceptual enrichment”

[14] Karen O’Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, Graciela Gonzalez, “Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions”, Arizona State University, Tempe, AZ Regis University, Denver, CO

[15] Jitendra Jonnagaddala, Toni Rose Jue , Hong-jie Dai “binary classification of twitter posts for adverse drug reactions”

[16] The jamovi project (2019). *jamovi*. (Version 1.0) [Computer Software]. Retrieved from https://www.jamovi.org.

[17] R Core Team (2018). *R: A Language and envionment for statistical computing*. [Computer software]. Retrieved from https://cran.r-project.org/.

[18] Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. [R package]. Retrieved from https://cran.r-project.org/package=psych.

[19] Rosseel, Y., et al. (2018). *lavaan: Latent Variable Analysis*. [R package]. Retrieved from https://cran.r-project.org/package=lavaan.

[20] K- Means Clustering Algorithm: How It Works: Analysis & Implementation https://www.educba.com/k-means-clustering-algorithm/

[21] “ICDData10.com” License ICD10Dta 2018 https://www.icd10data.com/search?s=drug+induces . [Accessed: 16-May-2020].

[22] Kathy Lee, Ashequl Qadir, Sadid A. Hasan, Vivek Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. 2017. Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. In Proceedings of the 26th International Conference on World Wide Web (WWW ’17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 705–714. DOI:https://doi.org/10.1145/3038912.3052671

[23] P Tafti A, Badger J, LaRose E, Shirzadi E, Mahnke A, Mayer J, Ye Z, Page D, Peissig P

Adverse Drug Event Discovery Using Biomedical Literature: A Big Data Neural Network Adventure

Appendix

```
[A]
import tweepy
import csv

# Fill the X's with the credentials obtained by
# following the above mentioned procedure.
consumer_key = "qkBPh6AyQdEZ1HbMZyJlCWYjt"
consumer_secret = "FileFXfTA6QnLFiRgmGvkdtnf8JZ2NLlKa9jpl6wKdhB0SGEbz"
access_key = "191424283-WKGqUOZNYQslqqtjK5ryrx3hwRBwUFOn8BirdIMi"
access_secret = "VJcFczwjYpKyHjKRQQybpMw7prL1eTRqzntOyN89uRzqN"



# Function to extract tweets
def get_tweets(username):
    # OAuth process, using the keys and tokens
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_key, access_secret)

    # Creation of the actual interface, using authentication
    api = tweepy.API(auth)

    results = api.search(username, count=500)

    csvfile = open('input.csv', 'wb')
    csvwriter = csv.writer(csvfile)

    for item in results:
    print(results)
    # write out the user, the tweet and their follower count into a
file
    # the unicode bits are required to write non ASCII language bits
into the file
    csvwriter.writerow([unicode(item.user.screen_name).encode("utf-
8"), unicode(item.text).encode("utf-8"),
                           item.user.followers_count])
    # time.sleep(1)



# Driver code
```

```
if __name__ == '__main__':
    # Here goes the twitter handle for the user
    # whose tweets are to be extracted.
    get_tweets("mucinex")
    get_tweets("#dettol")
    # get_tweets("nurofen")
    # get_tweets("@VeetIndia")
    # get_tweets("@clearasil")
    # get_tweets("@durex")
    # get_tweets("#dettol")
    # get_tweets("@MerckCanada")
    # get_tweets("@KnowHPV")
    # get_tweets("@KnowHPVForAll")
    # get_tweets("@Merck")
    # get_tweets("@MerckCanada")
    # get_tweets("@MerckCanada_FR")
    # get_tweets("@MerckEngage")
    # get_tweets("@MerckforMothers")
    # get_tweets("@MerckIMInspired")
    # get_tweets("@MerckManualHome")
    # get_tweets("@MerckManualPet")
    # get_tweets("@MerckManualPro")
    # get_tweets("@MerckSupport")
    # get_tweets("@MerckVetManual")
    # get_tweets("@MSDBelgium")
    # get_tweets("@msdcareers")
    # get_tweets("@MSDCyprus")
    # get_tweets("@MSDCzech")
    # get_tweets("@MSDDanmark")
    # get_tweets("@msdegypt")
    # get_tweets("@MSDEspana")
    # get_tweets("@MSDFinland")
    # get_tweets("@MSDFrance")
    # get_tweets("@MSDGreece")
    # get_tweets("@MSDinnofactory")
    # get_tweets("@MSDintheUK")
    # get_tweets("@MSDInvents")
    # get_tweets("@MSDKesfediyor")
    # get_tweets("@MSDLatAm")
    # get_tweets("@MSDLebanon")
    # get_tweets("@MSDManualHome")
    # get_tweets("@MSDManualPro")
    # get_tweets("@MSDNederland")
    # get_tweets("@MSDNorge")
    # get_tweets("@MSDOncologie")
    # get_tweets("@MsdPolska")
    # get_tweets("@msdsalute")
    # get_tweets("@MSDSerbia")
    # get_tweets("@MsdSlovensko")
    # get_tweets("@MSDSverige")
    # get_tweets("@MSD_Aus")
    # get_tweets("@MSD_Austria")
    # get_tweets("@MSD_Deutschland")
    # get_tweets("@msd_es")
    # get_tweets("@MSD_JAPAN")
    # get_tweets("@MSD_jordan")
    # get_tweets("@MSDBelgium")
    # get_tweets("@msd_portugal")
    # get_tweets("@MSD_Romania")
    # get_tweets("@MSD_Slovenia")
    # get_tweets("@MSD_SouthAfrica")
    # get_tweets("@MSD_Switzerland")
    # get_tweets("@MyMSDBeLux")
    # get_tweets("@OnlyMomCan")
    # get_tweets("@vreehealth_it")
    # get_tweets("g.zarez")
    # get_tweets("msdgcc")
```