Natural Language Processing Analysis of Reddit Data to Identify Substances Used in Opioid Withdrawal Treatment

Kaivil Brahmbhatt, *1153929* and Prithvisinh Jhala, *1137435*

Abstract

Opioids are natural or synthetic compounds that are used in health care settings to relieve pain, but they are also manufactured and taken for non-medical purposes. There may be withdrawal symptoms when an opiate user stop using them. To escape these effects, people frequently continue to abuse opioids. The purpose of this study is to develop an approach for identifying the medications/remedies that opioid users on the social media site (Reddit) take to manage withdrawal symptoms. Many of the withdrawal treatments mentioned by Reddit users have either been scientifically validated or may be of assistance. This implies that the method used in this project is a reliable tool for examining the self-treatment practises of an online community of opioid users.

I. Introduction

The main cause of ongoing opioid abuse and relapse in those who try to quit is withdrawal symptoms. Opioid withdrawal symptoms can be quite uncomfortable and persist within a week or more. Body aches, diarrhoea, nausea, vomiting, excessive perspiration, sleeplessness, and appetite loss are among the symptoms that are frequently present. Many individuals who use opioids to treat their symptoms relapse. A key focus of opioid therapy is the medical management of withdrawal; the National Institute on Drug Abuse has ranked the development of novel therapies for opioid use disorders as one of its top priorities [1].

Treatments for opioid withdrawal symptoms vary across clinicians. At this moment, tapering opioid agonists like methadone and buprenorphine—during which withdrawal symptoms may occur—is the most popular method of treating opioid use disorder. Standard therapies such loperamide for diarrhoea, ibuprofen for body pains, and ondansetron for vomiting are frequently used to treat these symptoms. As the first nonopioid drug designed exclusively to treat withdrawal symptoms, Luce Myra (lofexidine) was authorised by the US Food and Drug Administration (FDA) in 2018 [2]. In addition, some medical professionals prescribe off-label drugs (such baclofen and clonidine) to manage withdrawal. These drugs are frequently used to treat certain symptoms like nausea, diarrhoea, or body pains.

While a few opioid users seek professional assistance to lessen withdrawal symptoms, many do so through online discussion forums and blogs. To minimize their withdrawal symptoms, opioid addicts are actively trying alternative therapies. These treatments include over-the-counter drugs like loperamide for diarrhoea, more experimental drugs like supplements (such vitamins and herbs), and other techniques like yoga, acupuncture, and meditation [3]. These alternative therapies include some contentious ones (e.g., the use of the opioid-containing food supplement kratom). There is little knowledge on how opioid user's self-treat.

Machine learning techniques may be used to examine social media, which provides unique insights into the millions of online conversations concerning withdrawal treatments. Web-based discussion about opioid usage is popular due to the widespread engagement of the middle class in the opioid crisis, the popularity of social media, and the accessibility of smartphone devices. Twitter, smaller personal blogs, support groups, treatment facilities, and web-based forums like Reddit and Blue light are just a few of the sources that offer searchable and analysable data ideal for study. Discussion forums include an unmatched quantity of

knowledge regarding drug use and drug recovery; it is impossible to get such comprehensive information on drug use and recovery techniques anywhere else [4]. Recent research has examined forum data pertaining to buprenorphine, marijuana, social media, opioid recovery, and new drug use patterns. Others have demonstrated that online discussion of opioids relates with important surveillance data, such synthetic opioid mortality rates, and may be utilised as a leading indication [5]. Although research have started to leverage various sources, this data is still underutilised. There have been no evaluations of drugs that are used to treat withdrawal symptoms.

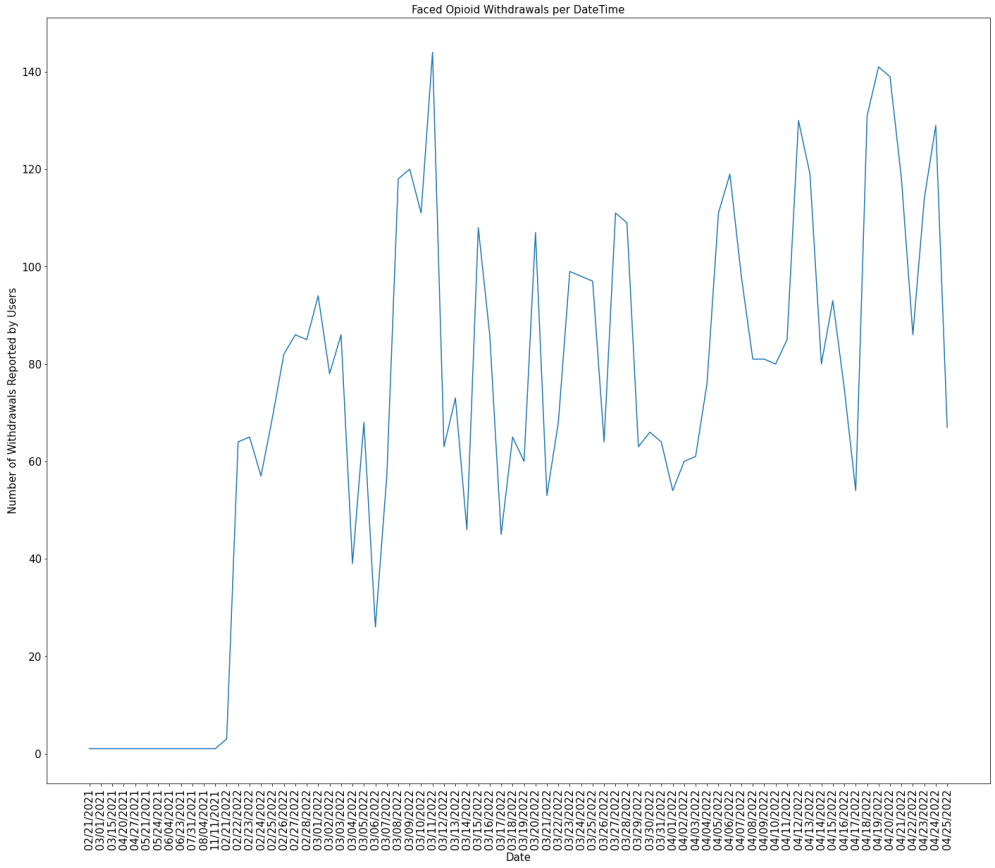


Chart 1: Line Chart of Number of Withdrawals Reported by Users per Date

II. OBJECTIVE

This study has two goals: (1) to verify a technique for investigating these self-treatment behaviours using social media (Reddit postings), and (2) to get a deeper understanding of these practises, including what is being employed and the results of such self-help. We point out that a tested system that analyses Reddit postings to comprehend problems like self-medication may be useful for a variety of medical and behavioural diseases.

The two Reddit discussion groups r/opiates and r/OpiatesRecovery are the main subjects of our work [6,7]. Both forums are devoted to having frank discussions regarding opioid use, frequently with the goal of assisting current and former users in getting well. Users frequently relate their experiences using conventional treatments and non-conventional therapies to lessen the symptoms of withdrawal as part of these talks.

III. LITERATURE REVIEW

The literature review has aided us in gaining a better knowledge of the opioid crisis in North America. The research that was published are quite instructive, and we sought to get as much information as possible from those. Machine learning and natural language processing approaches have shown to be quite effective in the past. We are attempting to better the work of researchers in the same field.

Natural language processing (NLP) discoveries in recent years have made it possible for researchers to recognise and extract information from massive quantities of text, including helpful data on prospective treatments. Numerous facets of the opioid crisis and other issues in the realms of public health and healthcare have been addressed using NLP [8]. The article" The Complete Practical Guide to Topic Modelling" by K. Yadav, aids data labelling requirements, since the themes developed in this stage may be applied to each set of similar documents [9]. The authors of the article "Drug side effect extraction from clinical narratives of psychiatry and psychology patients" devised a rule-based system for detecting the association between medications administered to psychiatry and psychology patients and their physician asserted adverse effects utilising NLP techniques in this study. Separately, they developed a system for extracting phrases that may contain adverse effects and medications that combines rule-based and machine learning approaches. This method was utilised to find as many adverse effect incidences in clinical notes as feasible [10].

In the study," Big data and predictive modelling for the opioid crisis: existing research and future potential", the potential for predictive analytics and routinely gathered administrative data to minimize overdose among patients with opioid use disorder is examined. To begin, they summarise global trends in opioid use and overdoses to identify gaps and the need for better interventions. Secondly, we have learned how big data has been used in opioid overdose research to date to understand how these resources have been employed. Additionally, the potential for predictive modelling, including machine learning, to prevent and monitor opioid overdoses should be considered. Finally, they have addressed the possible hurdles and hazards associated with using big data and machine learning to reduce opioid-related mortality and other harms [11]. The goal of the study," Detection of suicidality among opioid users on reddit: Machine learning-based approach" was to apply machine learning to extract postings indicating suicidality among opioid users on Reddit. The models' success is dependent on the quality of the data, and the findings will help us better understand the motivations of these users, bringing fresh insights into those affected by the opioid epidemic [12]. In the work" Automating the generation of lexical patterns for processing free text in clinical documents", the LDA approach was used as an algorithm to create topic modelling, each topic similarity, and visualisation of topic clusters from twitter data, which resulted in as many as four subjects (Economic, Military, Sports, and Technology) in Indonesian, each with a distinct number of tweets. In each topic extraction, topic modelling, producing index words that are in each topic cluster, and computer visualisation in the subject, the LDA approach employed in the processing of twitter data is effectively carried out and performs ideally. The LDA result outperforms the LSI technique in Topic Modelling in the process of word indexing in Sport themes with 1260 tweets, with an accuracy of 98 percent [13].

IV. PROPOSED METHODS

A. Data Collection

Reddit is a public social media platform made up of groups called subreddits that characterize material according to preferences. A post on Reddit is a submission. A post can be a discussion-starting passage of text or a link to another website. The latter is referred to as a selfpost, while the content is referred to as selftext. Other Reddit users may leave threaded comments on a post. Reddit submissions have three potential text sources for analysis: the post's title, its selftext (if it's a selfpost), and its threaded comments. We only focused on the comments that appear as discussions on a post and did not use text from post titles or selftext for analysis because our objective was to detect mentions of our entities within longer-form text and post titles and selftext frequently contain short phrases or incomplete sentences.

We conducted a thorough research into the names of the opioids used by the online community during this project's first phase in order to create a variety of generic names and its most prominent slang terms. These opioids oversaw the majority of opioid overdose incidents and whose cessation led to withdrawal symptoms in users. For the symptoms of opioid withdrawal, doctors employ a range of medications. The most popular medications for treating opioid use disorder right now are opioid agonists like methadone and buprenorphine. These medications are frequently tapered, and withdrawal symptoms may occur during this period. For example, loperamide for diarrhea, ibuprofen for body pains, and ondansetron for nausea are common therapies for these symptoms. The first non-opioid medicine authorised by the US Food and Drug Administration (FDA) to treat withdrawal symptoms is called Luce Myra (Lofexidine). For the abstract of the particular task, with the help of Web Scrapping we outperformed the task of data gathering from internet [14,15]. Furthermore, with this phase, the extracted names were than used to extract information by automatic link creation and web scrapping data from that link itself. Here, with the use of Beautiful Soup only we faced a problem on data extraction from few of the web pages, as those required Age Verification, so with the Beautiful Soup just being ideal and static in gathering data from the site, we brought in something more advance and dynamic, Selenium, it supports interacting with dynamic pages and content, hence with the Button Click accessibility for the Age Verification we can access the page source, and hence the comments and subreddit text, as shown in below figure(1).

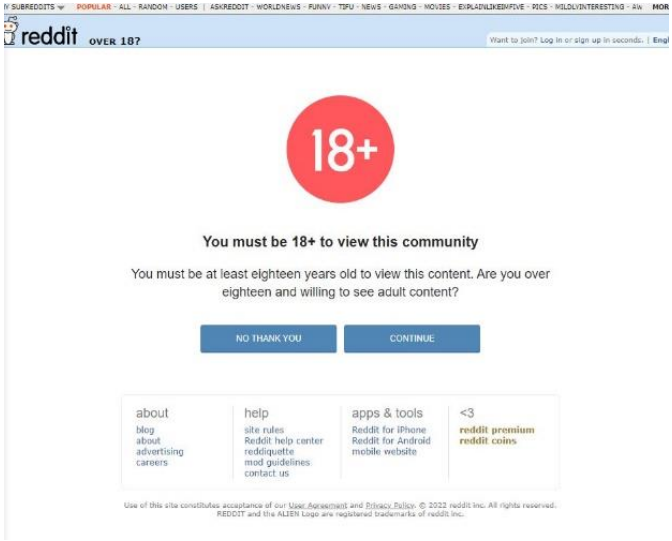


Figure 1: Age Restriction challenge

Moving on, it was discovered that there is more irrelevant information when the user is speaking entirely in slang or when the generic name or side effects are not mentioned in the remark or subreddit. So additionally, we employed a parallel approach for obtaining data from r/OpioidRecovery subreddit that was more appropriate for our project, Consequently, given the project's inclination, our main content sources were the r/opiates and r/OpiatesRecovery subreddits, both of which are related to recovery from opioid misuse and withdrawal side-effects.

Table 1. Summary of comments from r/opiates and r/OpiatesRecovery.

Item	Subreddit r/OpiatesRecovery
First comment, date; time	2021-02-21T03:13:29+00:00
Last comment, date; time	2022-04-25T18:20:19+00:00
Count	5410

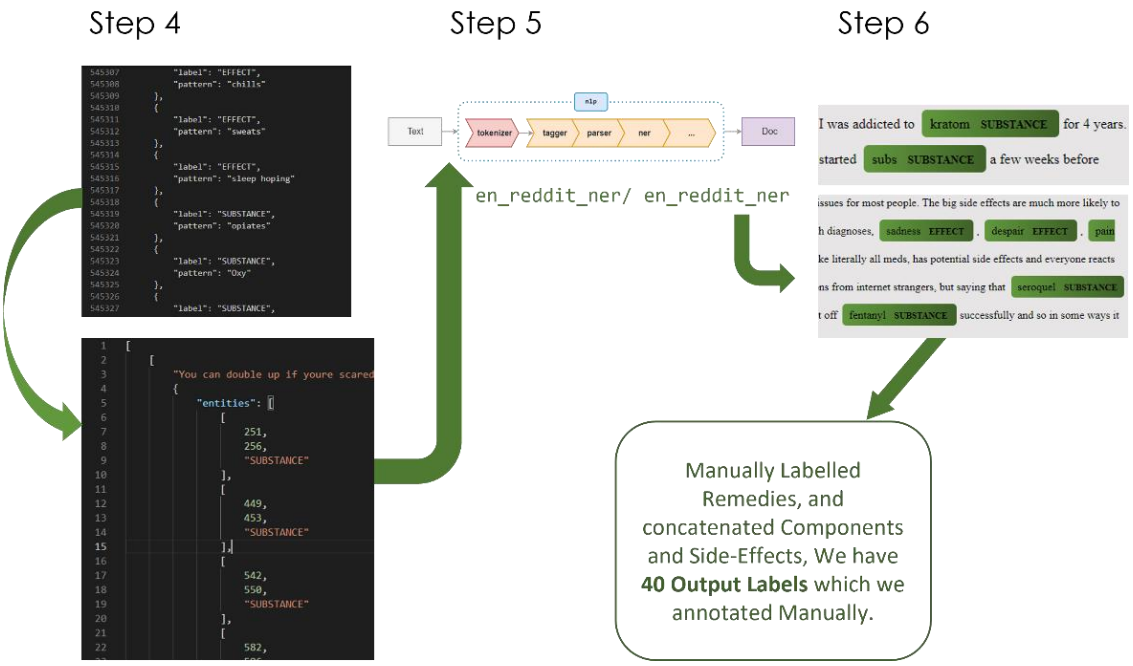


Figure 2: Dataset Processing and NER training.

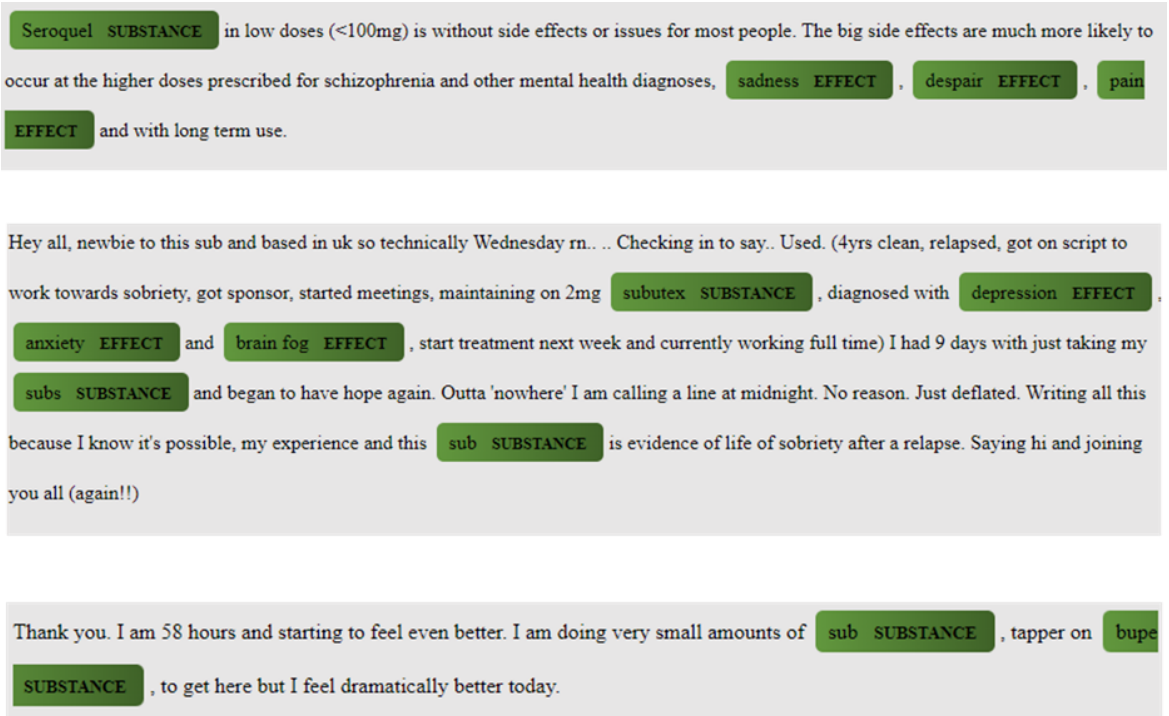


Figure 3: Example of entity recognition on general comments using Displacy library an inbuilt of spacy. [17]

B. NER Model

In order to explore Components-consumed and their withdrawal or side-effects, we structured our NLP work as a NER issue. We sought to distinguish between two potential entity types: Substances and Effects.

- Substance – drug, a substance that the users are consuming, and which is inducing side effects or tapering which is giving minor moderate or severe side effects.

- Effects – There are two types of effects related to the way of intake, positive and negative effects, one caused due to tapering and the other due to the consumption.

The 5410 classified comments that made up our training data set were created via an iterative data labelling and model training method. We used the standard spaCy (version 3.4.0; Explosion AI) [16] parameters for NER models to train our final NER model: 30 epochs with a dropout of 0.2 and compounding batch. We used 80% of the data to train the model, and 20% of the hold-out set was used to assess the model's performance. For accurate entity matches, precision, recall, and F1 scores were computed.

With the help of manual annotation and regular expressions, patterns were formed from the generic name table and using those patterns on the comments of all the reddit and twitter data we created our training set, figure (2). Examples are shown in figure (3).

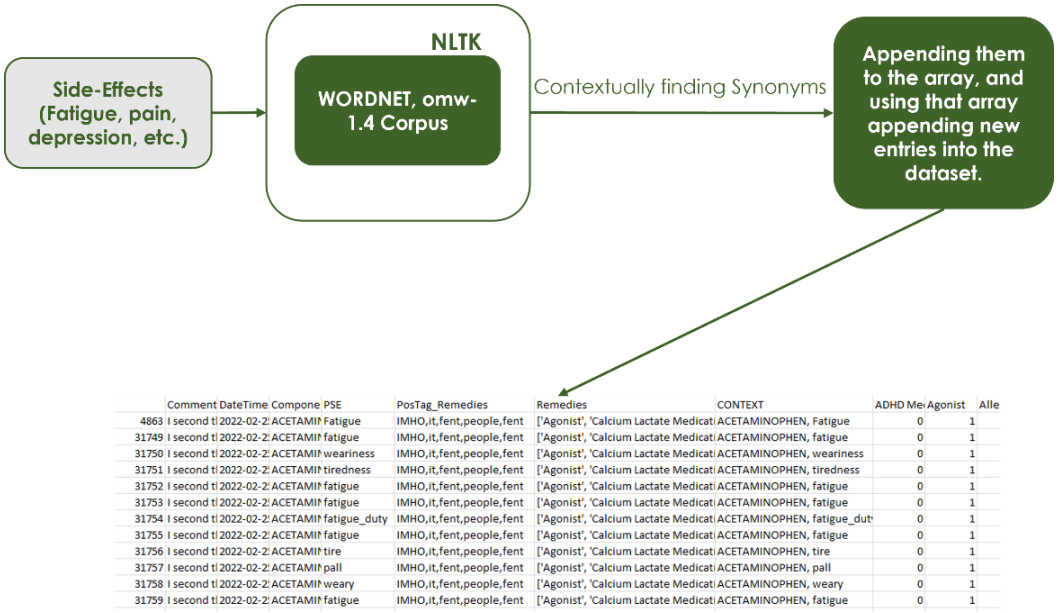
For the current phase, we have also labelled 40 labels, manually constructed the generalization and specification, where few of the labels like Methadone, Naloxone and Kratom etc. were not grouped according to the affinity with the Pain medication groups, as there are few labels which are most used by the users, but the ones with just 1 -10 occurrences are grouped in according to the specifications.

C. Data Augmentation

In order to avoid overfitting and increase generalisation of deep neural network models, data augmentation techniques are frequently used, processing the inputs for the model training and validation further, for these phase there were many techniques into consideration, few of which are mentioned below,

a.) BERT Contextual augmentation, is a recently proposed technique that improves labelled phrases by randomly substituting words with more diversified replacements suggested by the language model. BERT shows that a deep bidirectional language model is more effective than a shallow concatenation of a forward and backward model or a unidirectional language model. By adding a new conditional masked language model job, we convert BERT to conditional BERT. Contextual augmentation can be improved by using the well-trained conditional BERT. Studies on six distinct text classification tasks demonstrate that our strategy may be simply applied to classifiers using convolutional or recurrent neural networks to achieve a noticeable increase.[18]

b.) Back Translation for context-aware neural machine translation, some terms that are absent or confusing in the source languages must be added or specialised in the target languages. We require circumstances that go beyond a single sentence in order to interpret such ambiguous utterances, and we have thus far investigated context-aware neural machine translation (NMT). To train precise context aware NMT models, however, a sizable number of parallel corpora is not readily accessible. In this work, we first create massive pseudo-parallel corpora by back-translating monolingual data, and then we examine its influence on context-aware NMT models' translation accuracy. To show the significant impact of data augmentation for context-aware NMT models, we tested context-aware NMT models trained with tiny parallel corpora and the large-scale pseudo parallel corpora on English-Japanese and English-French datasets.[19]



	Comment	DateTime	Compone	PSE	PosTag_Remedies	Remedies	CONTEXT	ADHD Me	Agonist	Alle
4863	I second t	2022-02-2	ACETAMI	Fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, Fatigue	0	1	
31749	I second t	2022-02-2	ACETAMI	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, fatigue	0	1	
31750	I second t	2022-02-2	ACETAMI	weariness	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, weariness	0	1	
31751	I second t	2022-02-2	ACETAMI	tiredness	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, tiredness	0	1	
31752	I second t	2022-02-2	ACETAMI	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, fatigue	0	1	
31753	I second t	2022-02-2	ACETAMI	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, fatigue	0	1	
31754	I second t	2022-02-2	ACETAMI	fatigue_duty	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, fatigue_dut	0	1	
31755	I second t	2022-02-2	ACETAMI	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, fatigue	0	1	
31756	I second t	2022-02-2	ACETAMI	tire	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, tire	0	1	
31757	I second t	2022-02-2	ACETAMI	pall	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, pall	0	1	
31758	I second t	2022-02-2	ACETAMI	weary	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, weary	0	1	
31759	I second t	2022-02-2	ACETAMI	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calcium Lactate Medicat	ACETAMINOPHEN, fatigue	0	1	

Figure 4: Data Augmentation of Train Data.

c.) Synonym Replacement technique for Data augmentation, with vary in words we found that there is diversity in the number of synonyms it may consist, as shown in figure (4) using the Wordnet and omw-1.4 finding the semantic relationship between words, and grouping them into Synsets which contains metadata of the words passed to it for finding the synonyms, from the dataset the entry copy is been made and appended to the same dataset but with the replacement of the specific words from the side effects, as inputs would be just the concatenation of the Side-Effects and the Generic Component Name, as the form of Context to the model for training and validation. Increasing the size of the Dataset as shown in the table (3) below,

Label Name	Label Count
ADHD Medication	337
Agonist	2911
Allergic Reaction Medication	832
Amino Acid Medication	112
Antiepileptic Medication	485
Antihistamine Medication	169
Anxiety Medication	6431
Barbiturates	1527
Benzodiazepines	1385
Blood Pressure Medication	63
Calcium Lactate Medication	4746

Fever Medication	4284
Gateway Drugs	570
High blood Pressure Medication	186
Insomnia Medication	1691
Mild/Moderate Pain Medication	1503
Moderate/Severe Pain Medication	1750
Muscle Pain Medication	12357
NSAIDS	4759
Ondansetron	459
Opioid Analgesic Medication	1821
Opioid Overdose Medication	1059
Rehab and Detox Medication	2534
Ropinirole Medication	234
Stomach Medication	638
Vitamin deficiency Medication	1256
alkaloids	274
antacids	1971
caffeine	574
kratom	3353
methadone	320
naltrexone	801
probiotics	277
psychedelics	189
quetiapine	416
sour throat Medication	622
suboxone	3398

Table 2. Label Names ad their corresponding Counts

Table 3. Summary of Augmented Train-Dataset.	
Item	Main-Dataset
Length before Data Augmentation of Train Data	4868
Length after Data Augmentation of Train Data:	26575

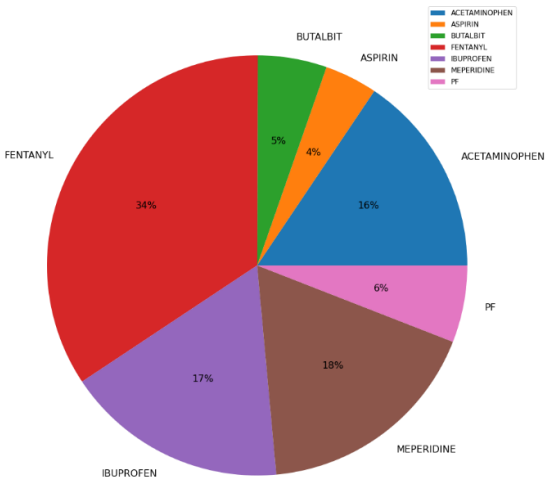


Chart 2: Category of labels with a count greater than 500 in the Dataset.

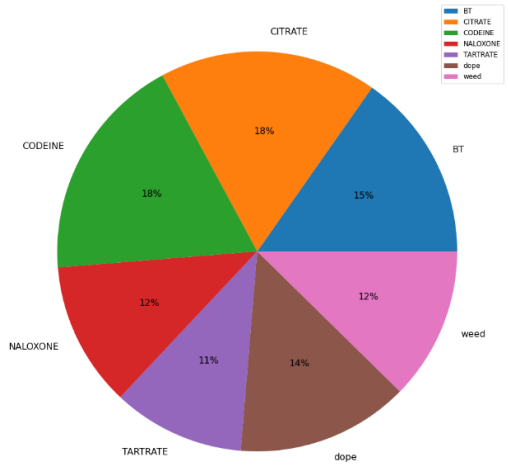


Chart 3: Category of labels with a count less than 500 and more than 200 in the Dataset.

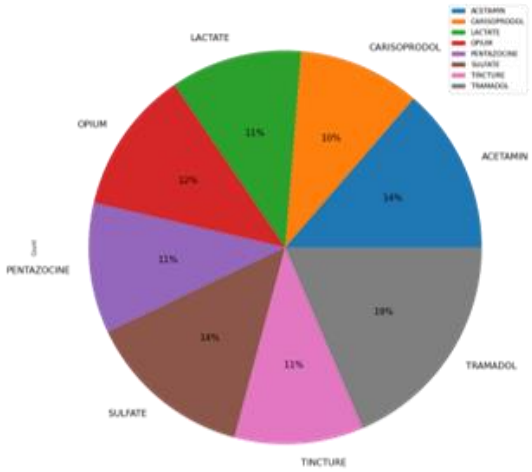


Chart 4: Category of labels with a count Less than 200 and greater than 100 in the Dataset.

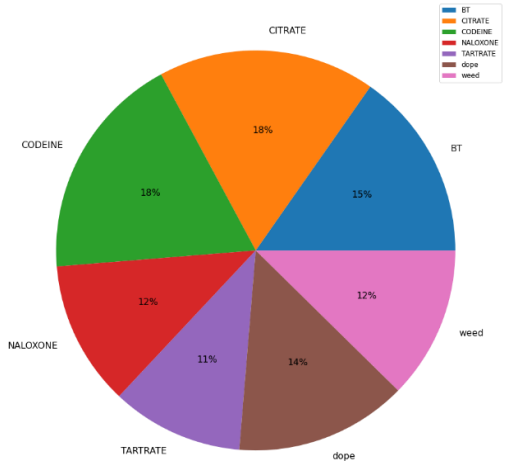


Chart 5: Category of labels with a count less than 100 in the Dataset.

i. FATIGUE

Comment	DateTime	Compone	PSE	PosTag_Remedies	Remedies	CONTEXT
I second tl	2022-02-2	ACETAMIN	Fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calciu	ACETAMINOPHEN, Fatigue
I second tl	2022-02-2	ACETAMIN	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calciu	ACETAMINOPHEN, fatigue
I second tl	2022-02-2	ACETAMIN	weariness	IMHO,it,fent,people,fent	['Agonist', 'Calciu	ACETAMINOPHEN, weariness
I second tl	2022-02-2	ACETAMIN	tiredness	IMHO,it,fent,people,fent	['Agonist', 'Calciu	ACETAMINOPHEN, tiredness
I second tl	2022-02-2	ACETAMIN	fatigue	IMHO,it,fent,people,fent	['Agonist', 'Calciu	ACETAMINOPHEN, fatigue

ii. SHAKINESS

Comment	DateTime	Component	PSE	PosTag_Remedies	Remedies	CONTEXT	
Got my vi	2022-02-27	IBUPROFEN	shakiness	HBO,Harlem,God,d	['naltrexone', 'Opioid Ove	IBUPROFEN, shakiness	
Got my vi	2022-02-27	IBUPROFEN	shaking	HBO,Harlem,God,d	['naltrexone', 'Opioid Ove	IBUPROFEN, shaking	
Got my vi	2022-02-27	IBUPROFEN	trembling	HBO,Harlem,God,d	['naltrexone', 'Opioid Ove	IBUPROFEN, trembling	
Got my vi	2022-02-27	IBUPROFEN	quiver	HBO,Harlem,God,d	['naltrexone', 'Opioid Ove	IBUPROFEN, quiver	
Got my vi	2022-02-27	IBUPROFEN	quivering	HBO,Harlem,God,d	['naltrexone', 'Opioid Ove	IBUPROFEN, quivering	

iii. SEIZURE PAIN

Comment	DateTime	Componen	PSE	PosTag_Remedies	Remedies	CONTEXT
Good for y	2022-02-2	ACETAMIN	seizure	opiates,addiction,y	['Calcium Lact	ACETAMINOPHEN, seizure
Good for y	2022-02-2	ACETAMIN	ictus	opiates,addiction,y	['Calcium Lact	ACETAMINOPHEN, ictus
Good for y	2022-02-2	ACETAMIN	raptus	opiates,addiction,y	['Calcium Lact	ACETAMINOPHEN, raptus
Good for y	2022-02-2	ACETAMIN	capture	opiates,addiction,y	['Calcium Lact	ACETAMINOPHEN, capture
Good for y	2022-02-2	ACETAMIN	gaining_control	opiates,addiction,y	['Calcium Lact	ACETAMINOPHEN, gaining_contr

D. Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa)

BERT is a model that set several benchmarks for how effectively models can perform tasks involving language. The model's code was also made available for download, along with versions of the model that had previously been pre-trained on sizable datasets, shortly after the team's article detailing the model was published. This is a significant advance because it allows anybody creating a machine learning model that uses language processing to leverage this powerful component right now, saving the time, effort, knowledge, and resources required to train a language-processing model from start.[20]

With the use of BERT(Base) model available by the Transformer library, with configuration as follows, 110 Global Parameters, 12 encoders (12 attention heads) and 768 number of input ids (also number of hidden layer). In this case, we modify the pre-trained BERT model to better suit our categorization goal. Basically, we load the previously trained model, train the last layer, and then do the classification job.

The technique for tokenization includes the unique "CLS" and "SEP" tokens that BERT uses to denote the beginning and end of a phrase, respectively. Additionally, it adds the tokens "index" and "segment" to each input. Thus, this function handles the entire task of formatting input in accordance with the BERT. Before calculating the likelihood, using logits (score), which are unscaled, unprocessed numbers connected to a class. This implies that a logit is an output of a dense (fully connected) layer in terms of neural network design. A Sigmoid layer and the BCELoss are combined into a single class by the usage of nn.BCEWithLogitsLoss() loss. By integrating the operations into one layer, we are able to take use of the log-sum-exp method for numerical stability, making this version more stable numerically than one that uses a simple Sigmoid followed by a BCELoss [21].

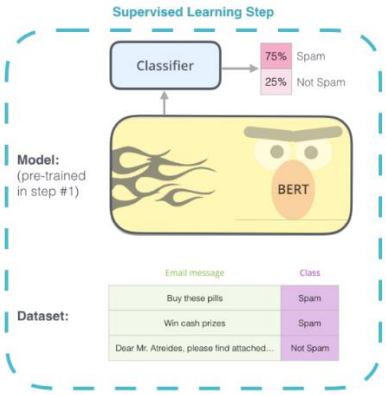


Figure 5: Supervised Training on a specific Task with the labeled dataset [20]

$$\ell(x, y) = \begin{cases} \text{mean}(L), & \text{if reduction} = \text{'mean'}; \\ \text{sum}(L), & \text{if reduction} = \text{'sum'}. \end{cases}$$

Equation 1: BCE with Logits Loss [21]

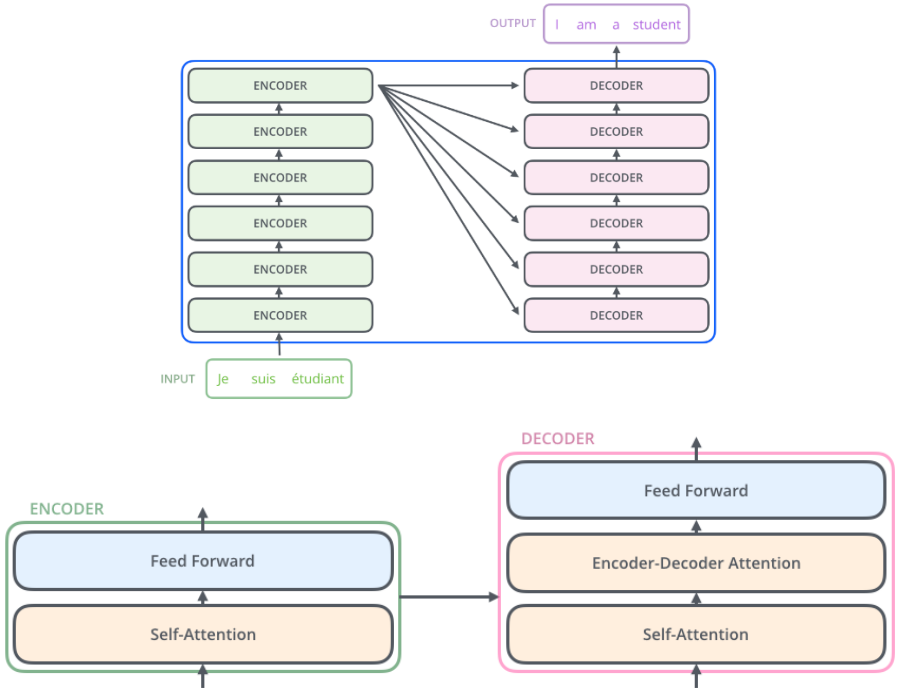


Figure 6: Encoder-Decoder Model [20]

The optimizer AdamW features a better weight decay implementation than the original Adam, for optimization purpose we have finetuned two main parameters of this optimizer

- Learning-Rate, means that the weights are adjusted more often with each iteration, which might make them achieve their ideal value more quickly but also cause them to miss it, currently we have used the default Learning rate that is 5e-5.
- To reduce the likelihood of overfitting, weight decay is a type of regularization, that to we have processed using the default values

The Major difference between RoBERTa and BERT is that its does not contain Next Sentence Prediction process inside it.

E. XLNet (generalized autoregressive language model)

[22] Pretraining methods based on denoising autoencoding, such BERT, perform better than those based on autoregressive language modelling because they can model bidirectional contexts. BERT neglects the dependence between the masked locations and experiences a pretrain-finetune mismatch since it relies on masking the input to corrupt it. With these advantages and disadvantages in mind, we suggest XLNet, a generalized autoregressive pretraining method that

- Makes it possible to learn bidirectional contexts by maximizing the expected likelihood over all variations of the factorization order and
- All Thanks to its autoregressive formulation, outperforms the drawbacks of BERT. Furthermore, Transformer-XL, a cutting-edge autoregressive model, is incorporated into pretraining by XLNet.

XLNet is a Bidirectional Transformer with Permutation based modelling as shown in figure (7), it does not predict just 15% (like BERT and RoBERTa) of the random data but it goes through each token in a permutation fashion. This helps the model to learn the context in a more generous and open way, but also with more computational need.

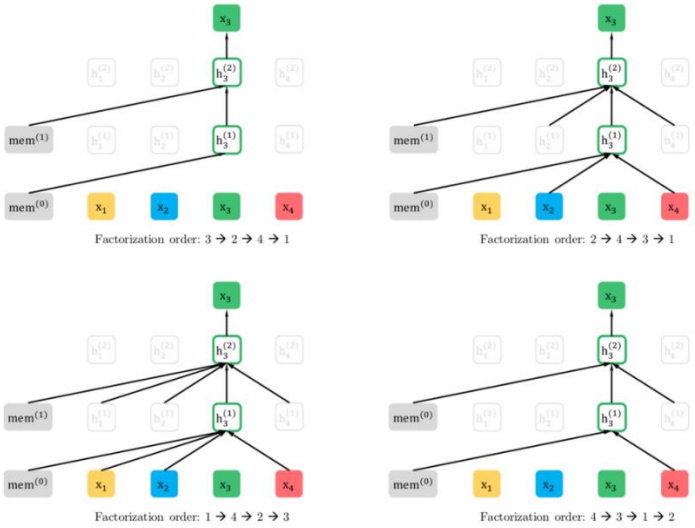


Figure 7: Example of permutation language modeling objective for predicting x_3 given the same input sequence x but with different factorization orders.[22]

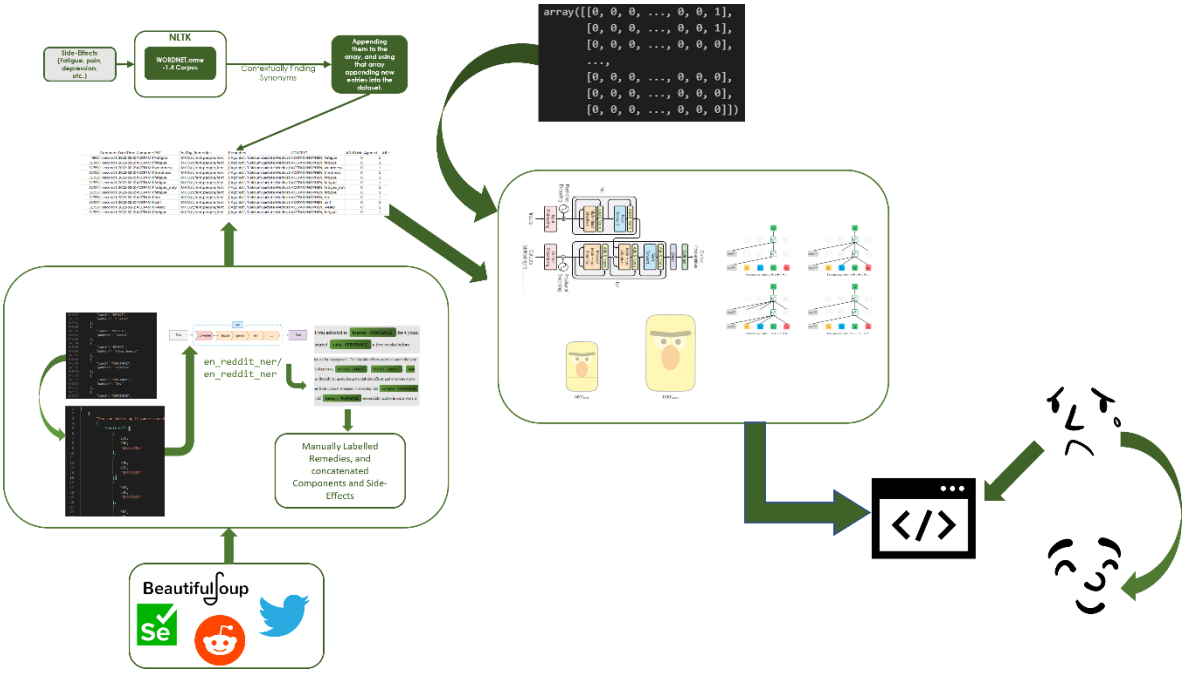


Figure 8: Flow Chart

F. Evaluation

Keeping the validation in mind we have divided the data from the main dataset into train and test with a ratio of 80% and 20%, and further divided the data into train and validation from the 80% Trainset, with the same ratio again. With the validation accuracy we are comparing the predicted logits with the true logits batch-wise in a iterative fashion, finding validation of each batch per epoch, and for loss again nn.BCEWithLogitsLoss() loss is used. Also evaluated traditional metrices to monitor system performance to predict remedies based on the label.

F1-Score, Precision and Recall are counted in two structures,

- The first experiments the F1-Score, Precision and Recall keeping the whole dataset validation in mind, evaluating global score on the Test Dataset
- And, the other one iterates through each label evaluating local label scores.

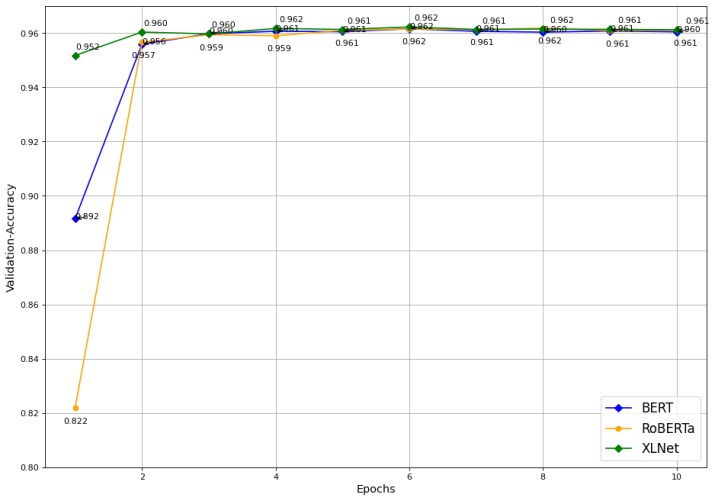


Chart 6: Validation Accuracy Comparison using Line Graph

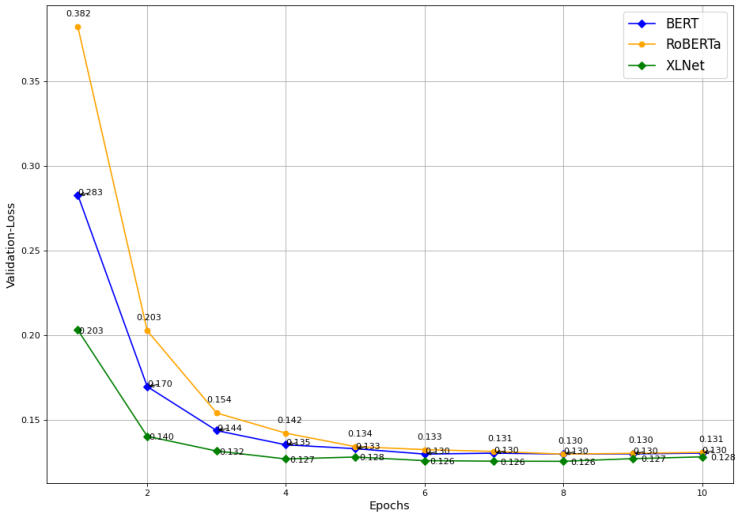


Chart 7: Validation Loss Comparison using Line Graph

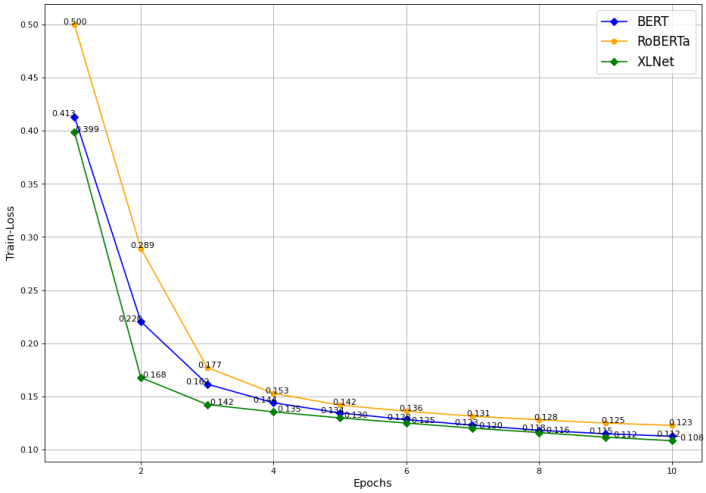


Chart 8: Train Loss Comparison using Line Graph

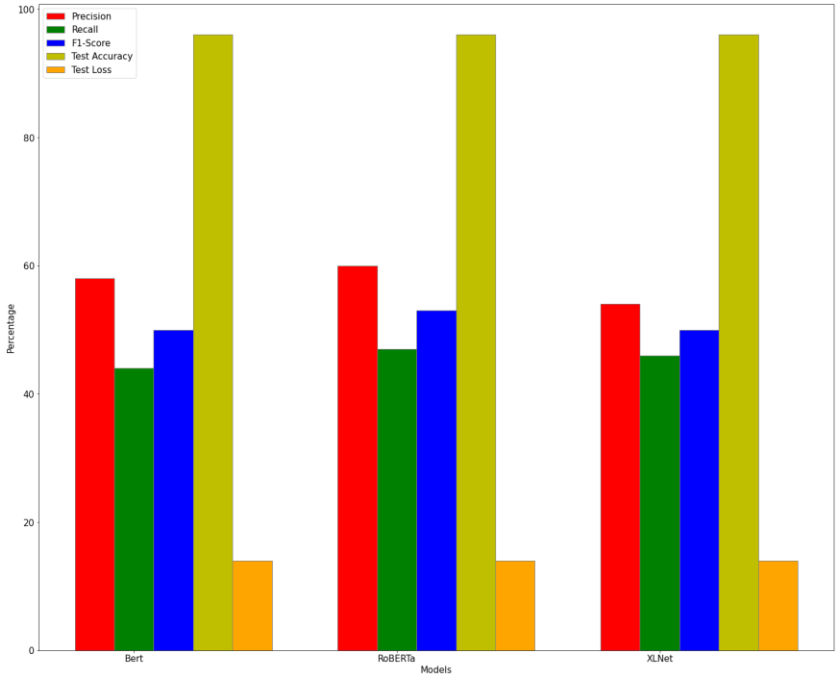


Chart 9: Model Validation of BERT, RoBERTa, and XLNet

V. LIMITATIONS AND FUTURE WORK

Our study has a number of drawbacks. The absence of causation in the relationships between substances and consequences is foremost. It is impossible to tell without reading the posts if the drug itself produced the impact or whether it was used to mitigate it. Furthermore, we haven't yet examined how effective the treatments were. The phrase "gabapentin treated my body pains," "I tried gabapentin for body aches, but it didn't work," or "gabapentin gave me body aches" are only a few examples of associations. The goal of this study, however, was to provide and establish the groundwork for future studies. Any subset of the relationships found in the data might be the subject of more thorough investigations.

In subsequent research, we intend to examine specific combinations of chemicals and effects using various NLP approaches in order to determine whether the link was beneficial.

We only considered connections that occurred inside a single phrase. When the effect and substance are described in separate phrases (for instance, "I feel awful bodily pains. Gabapentin is helpful for me while aspirin is ineffective. For the purposes of clarity and simplicity, we restricted the technique in this study to sentences. As a result of our success with single sentences, we will concentrate on integrating linkages spanning many phrases in our future work. Determining the boundaries of paragraphs and untangling the numerous links between various drugs and effects across sentences make this endeavour challenging. The ambiguity of postings' free-flowing content is a frequent problem in social media analysis.

The application of transformer models, such as Bidirectional Encoder Representations from Transformers, has considerably advanced the area of NLP since the data were gathered and the models were constructed. This has improved general NER jobs from an F1 of 86 in 2017 to >90 in 2019. The focus in the field has been shifting from model-centric approaches (e.g., hyperparameter tuning) to data-centric approaches (e.g., higher-quality labelled data), as, in many scenarios, more benefit comes from data than architecture. Large transformer models require significant specific computational resources to train and deploy compared to more traditional methods. In conclusion, our inability to employ state-of-the-art models is a constraint of our study. Future research will analyse possible gains from new model designs and increased data quality.

Substances, effects, and substance-effect pairings were grouped into categories throughout our expert review and validation processes based only on the terms' literal meanings, without a close examination of the post content itself. There may be variation in categories due to the differing perspectives and interpretations of various reviewers. We want to use the Withdrawal Remedy Explorer tool in future projects to get feedback and adjustments from other professionals.

In order to relate the extracted entities and their networks to other knowledge bases, we will think about creating a knowledge graph as our study creates networks of extracted knowledge.

Despite these restrictions, we were nevertheless able to find and confirm powerful signal in the information. The finding of effective withdrawal therapies motivates us to examine the data's more subtle nuances. The constraints of our study direct us toward our research's future phases.

VI. CONCLUSION

In this work, we verified a method for finding opioid withdrawal treatments using the online community Reddit. We created a process to extract chemicals and effects from unprocessed data, found significant correlations between prospective treatments and withdrawal symptoms, and confirmed these correlations. Our findings show that internet and online forum conversations might provide insight into how people manage withdrawal symptoms. This information might be used to spot potentially harmful new therapies, as well as public health issues.

REFERENCES

[1] Volkow ND, Collins FS. The role of science in addressing the opioid crisis. N Engl J Med 2017;377(4):391-394.

[2] Doughty B, Morgenson D, Brooks T. Lofexidine: a newly FDA-approved, nonopioid treatment for opioid withdrawal. Ann Pharmacother 2019;53(7):746-753.

[3] Alfonso III F. How a Reddit forum has become a lifeline to opioid addicts in the US. The Guardian. 2017. URL: https://www.theguardian.com/society/2017/jul/19/opioid-addiction-reddit-fentanyl-appalachia [accessed 2022-07-23]

[4] Chancellor S, Nitzburg G, Hu A, Zampieri F, De Choudhury M. Discovering alternative treatments for opioid use recovery using social media. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK p. 1-15.

[5] Bowen DA, O'Donnell J, Sumner SA. Increases in online posts about synthetic opioids preceding increases in synthetic opioid death rates: a retrospective observational study. J Gen Intern Med 2019;34(12):2702-2704

[6] Opiates. Reddit. URL: https://www.reddit.com/r/opiates/ [accessed 2022-06-20]

[7] You are not alone in this fight. Reddit. URL: https://www.reddit.com/r/OpiatesRecovery/ [accessed 2022-07-15]

[8] Chancellor S, Nitzburg G, Hu A, Zampieri F, De Choudhury M. Discovering alternative treatments for opioid use recovery using social media. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK p. 1-15.

[9] K. Yadav, "The Complete Practical Guide to Topic Modelling," Medium, 22-Jan-2022. [Online]. Available: https://towardsdatascience.com/topic-modelling-f51e5ebfb40a. [Accessed: 03-Mar-2022].

[10] S. Sohn, J.-P. A. Kocher, C. G. Chute, and G. K. Savova, "Drug side effect extraction from clinical narratives of psychiatry and psychology patients," OUP Academic, 21-Sep-2011. [Online]. Available: https://academic.oup.com/jamia/article/18/Supplement 1/i144/797245. [Accessed: 03-Mar-2022].

[11] C. Bharat, M. Hickman, S. Barbieri, and L. Degenhardt, "Big data and predictive modelling for the opioid crisis: existing research and future potential," The Lancet Digital Health, 2021. [Accessed: 03-Mar-2022].

[12] H. Yao, S. Rashidian, X. Dong, H. Duanmu, R. N. Rosenthal, and F. Wang, "Detection of suicidality among opioid users on reddit: Machine learning-based approach," Journal of medical Internet research, 27-Nov-2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7732714/. [Accessed: 03-Mar-2022].

[13] F. Meng and C. Morioka, "Automating the generation of lexical patterns for processing free text in clinical documents.," Journal of the American Medical Informatics Association: JAMIA, Sep-2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4986670/. [Accessed: 03-Mar-2022].

[14] D. Balsamo, P. Bajardi, A. Salomone, and R. Schifanella, "Patterns of routes of administration and drug tampering for nonmedical opioid consumption: Data Mining and content analysis of reddit discussions," Journal of Medical Internet Research, 04-Jan-2021. [Online]. Available: https://www.jmir.org/2021/1/e21212/.

[15] FDA, "FDA approves the first non-opioid treatment for management of opioid withdrawal symptoms in adults," U.S. Food and Drug Administration, 16-May-2018. [Online]. Available: https://www.fda.gov/news-events/press-announcements/fda-approves-first-non-opioid-treatment-management-opioid-withdrawal-symptoms-adults.

[16] "Linguistic features · spacy usage documentation," Linguistic Features. [Online]. Available: https://spacy.io/usage/linguistic-features#named-entities.

[17] "DisplaCy dependency visualizer · explosion," Explosion. [Online]. Available: https://explosion.ai/demos/displacy.

[18] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional BERT contextual augmentation," arXiv.org, 17-Dec-2018. [Online]. Available: https://arxiv.org/abs/1812.06705.

[19] A. Sugiyama and N. Yoshinaga, "Data augmentation using back-translation for context-aware neural machine translation," ACL Anthology, Nov-2019. [Online]. Available: https://aclanthology.org/D19-6504/.

[20] J. Alammar, "The illustrated Bert, Elmo, and Co. (how NLP cracked transfer learning)," The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) – Jay Alammar – Visualizing machine learning one concept at a time., 03-Dec-2019. [Online]. Available: https://jalammar.github.io/illustrated-bert/.

[21] "BCEWITHLOGITSLOSS," BCEWithLogitsLoss - PyTorch 1.12 documentation. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html.

[22] XLNet. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/xlnet.