Classifying Adverse Event Seriousness using NLP

Capstone

Gan Tze Ling
DSI-25

Problem Statement

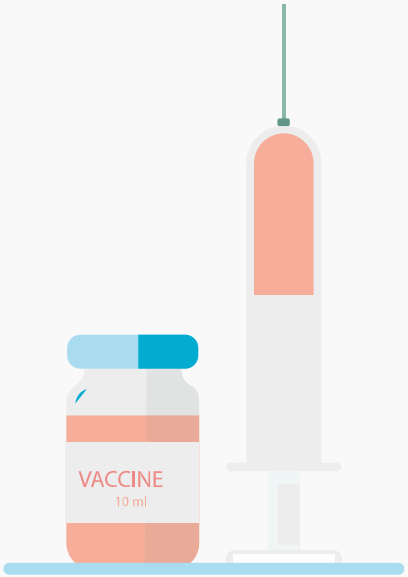
As a data scientist in a consultant firm to the **Health Authority** in Singapore, we have been tasked to create a model to **differentiate** between **serious and non-serious adverse events (AE)** using **Natural Language Processing (NLP)** from reports obtained from various sources.

The following models will be tested as potential candidates:

Logistic Regression
Naive Bayes - Multinomial
Random Forest Classifier
Ada Boost Classifier
Support Vector Machine (SVM)

A successful model is defined as having an **accuracy** and **F1 score** of at least 0.7.

TABLE OF CONTENTS



01

INTRODUCTION

Background
Dataset
Data Cleaning

BACKGROUND

DRUG SAFETY

After regulatory approval of a drug, the ongoing process of post-market surveillance ensures continued safety of the product

ADVERSE EVENT

Harmful or negative outcome that occurs when a patient has been provided with medical care or treatment

VAERS

Vaccine Adverse Events Reporting System (VAERS) is national warning system in the US to detect possible safety problem in US-licensed vaccines

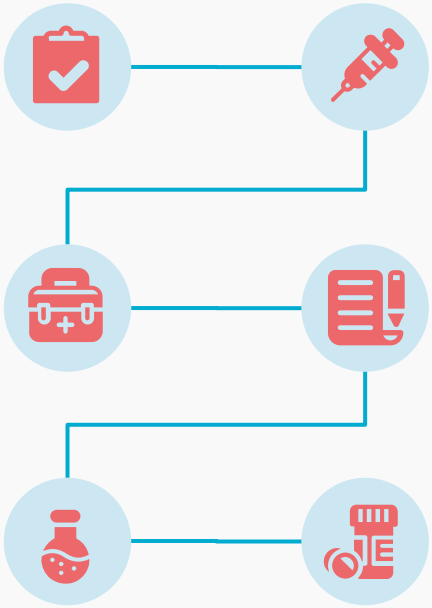
Reporting of an Adverse Event

1. Product approval

Efficacy and safety have been tested in clinical trials

2. Product Administration

Patient is given the drug or vaccine



3. Adverse Event

Patient could experience a side effect of the drug

4. Reporting

AE is reported by anyone (patient, family, HCP)

5. Signal Detection

If there is a high incidence of a particular AE, it could be detected as a signal for regulatory action

6. Action Taken

Regulatory authorities together with the company will determine what is the best course of action (e.g. Advisory, Recall)

INITIAL DATASET

- Taken from VAERS
- Data Collected in 2021 (up to Oct)
- From 3 CSV files

FEATURES

52

COLUMNS

830k

ROWS

DESCRIPTION

45 Non-text columns

7 Text Columns

DATA CLEANING

Removal of Duplicates

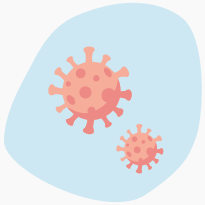
Imputing or Removal of Null Values

Cleaning of Text Columns with Regex

TYPES OF AE REPORT

NON-SERIOUS

Common non-serious AE include
headache, fever, nausea



SERIOUS

Generally more serious, falls into one of the serious criteria

SERIOUS CRITERIA

- Death
- Life-threatening
- Hospitalisation (initial or prolonged)
- Disability or Permanent Damage
- Congenital Anomaly or Birth Defect
- Other Serious (Important Medical Events (IME))

IME LIST

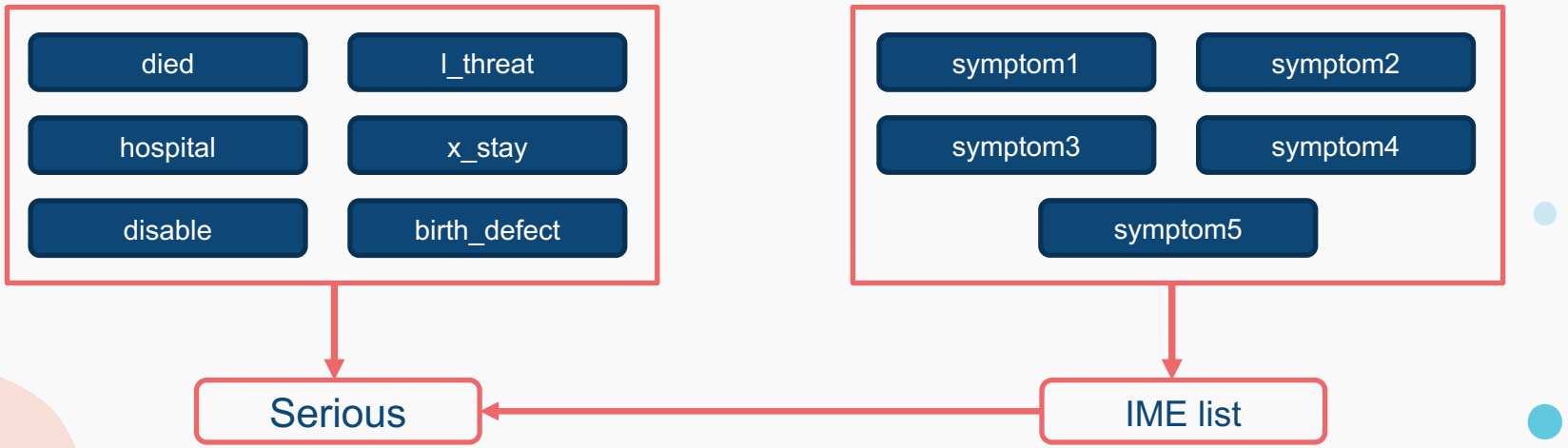
	MedDRA Code	PT Name	SOC Name	Comment	Added in 24.0	Primary SOC Change
0	10083258	Erythropoietin deficiency anaemia	Blood and lymphatic system disorders	Existing PT. Added after review by EVEWG.	X	NaN
1	10051778	Factor IX inhibition	Blood and lymphatic system disorders	Existing PT. Added after review by EVEWG.	X	NaN
2	10048619	Factor VIII inhibition	Blood and lymphatic system disorders	Existing PT. Added after review by EVEWG.	X	NaN
3	10058116	Nephrogenic anaemia	Blood and lymphatic system disorders	Existing PT. Added after review by EVEWG.	X	NaN
4	10068698	Familial hypocalciuric hypercalcaemia	Congenital, familial and genetic disorders	Existing PT. Added after review by EVEWG.	X	NaN

Creation of ‘Serious’ Column as Target Variable

Columns Representing Serious Criteria

Columns for Important Medical Events (IME)

The IME list can be found from MedDRA



DATASET AFTER CLEANING

FEATURES

32

COLUMNS

Merged into a single dataset

543k

ROWS

DESCRIPTION

28 Non-Text Columns

4 Text Columns

Target Feature: ‘serious’

02

EXPLORATORY DATA ANALYSIS

Pre-processing of Data
Data Visualisation

PRE-PROCESSING OF TEXT COLUMNS

REMOVE PUNCTUATIONS

Clean text of any punctuations

REMOVE STOPWORDS

Exclude words that are frequently occuring in all text

VECTORIZATION

Count Vectorizer
TF-IDF Vectorizer

TOKENIZATION

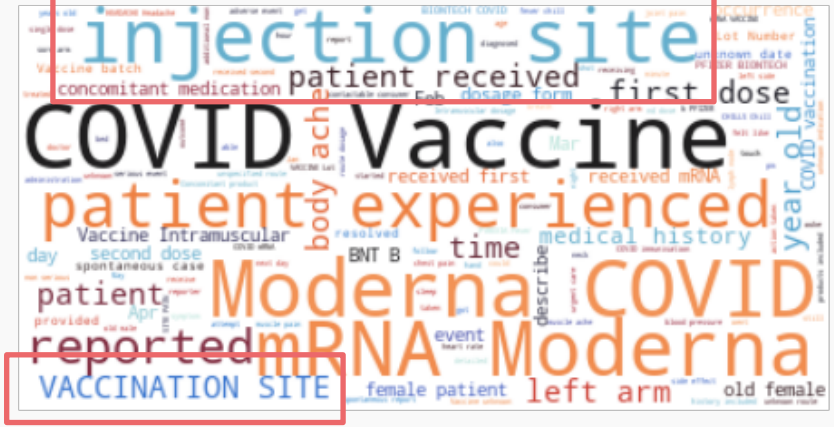
Splitting text into a list of words

LEMMATIZATION

Reducing all words to their base form

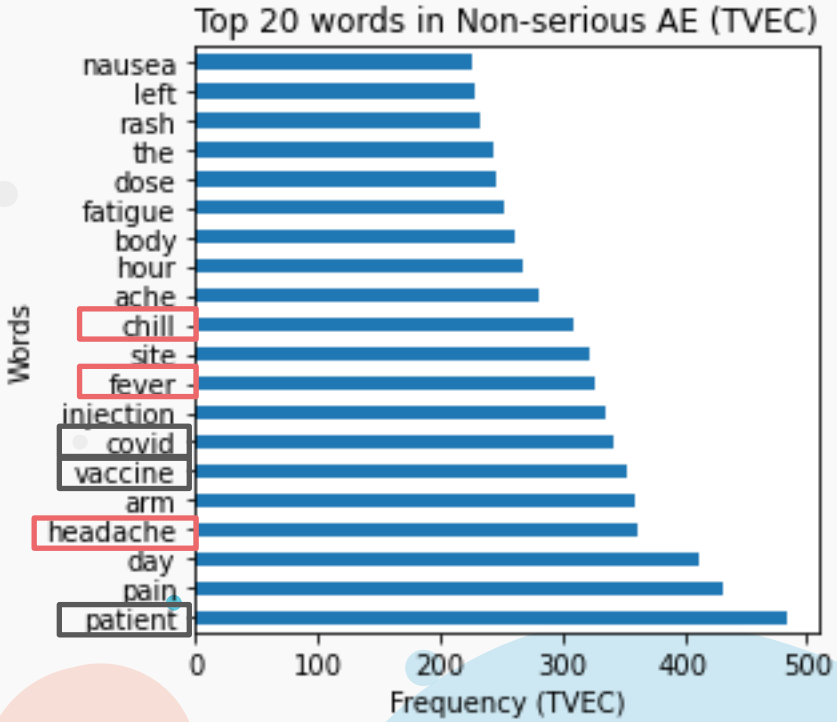
Word Cloud – symptom_text

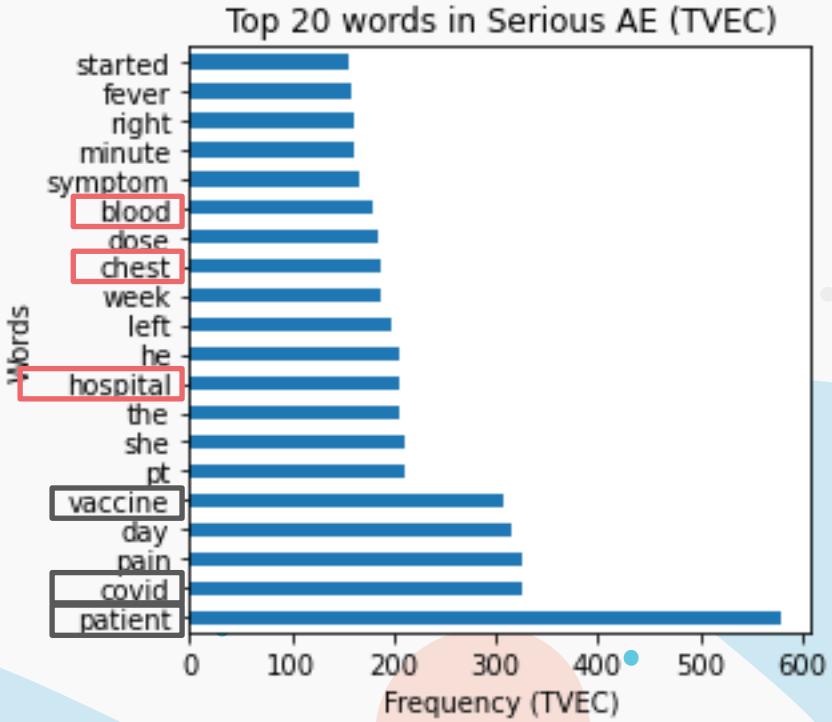
Word Cloud for Non-serious AE



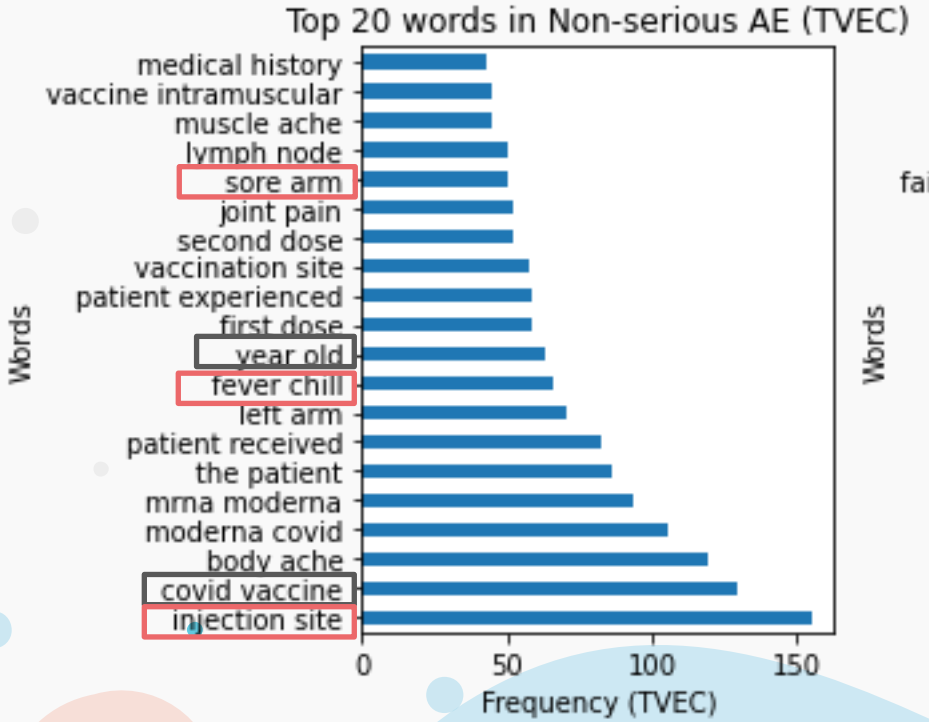
Word Cloud for Serious AE

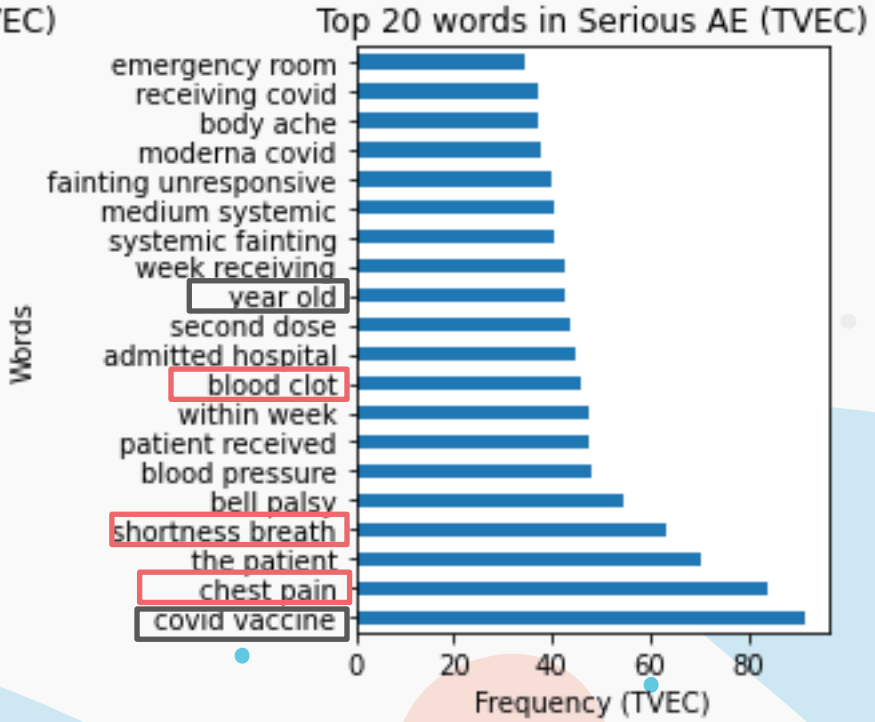
Unigram – symptom_text





Bigram – symptom_text



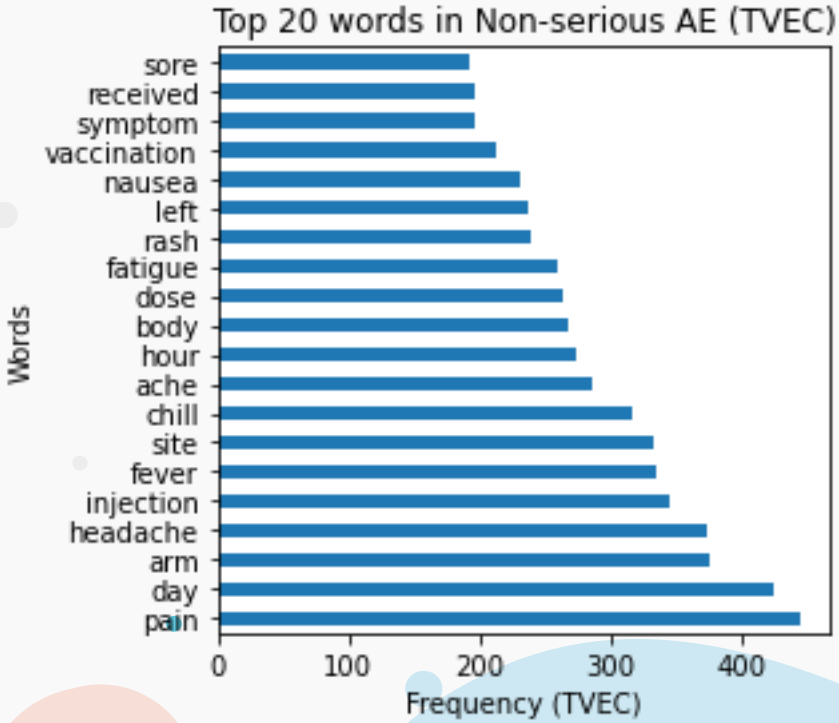


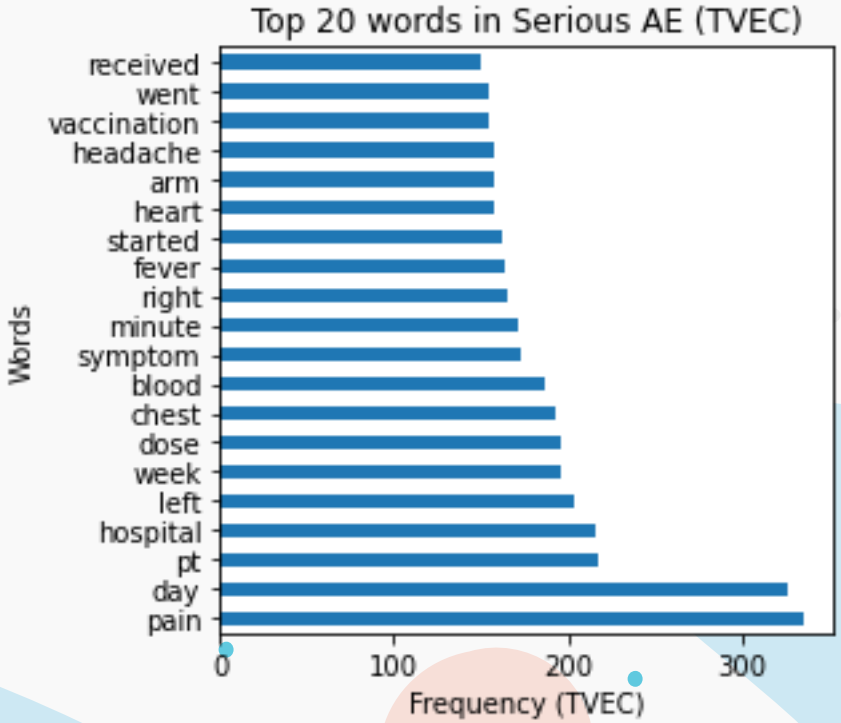
New Stopwords

- Words that do not have predictive power or are too repetitive were added to the list of new stopwords

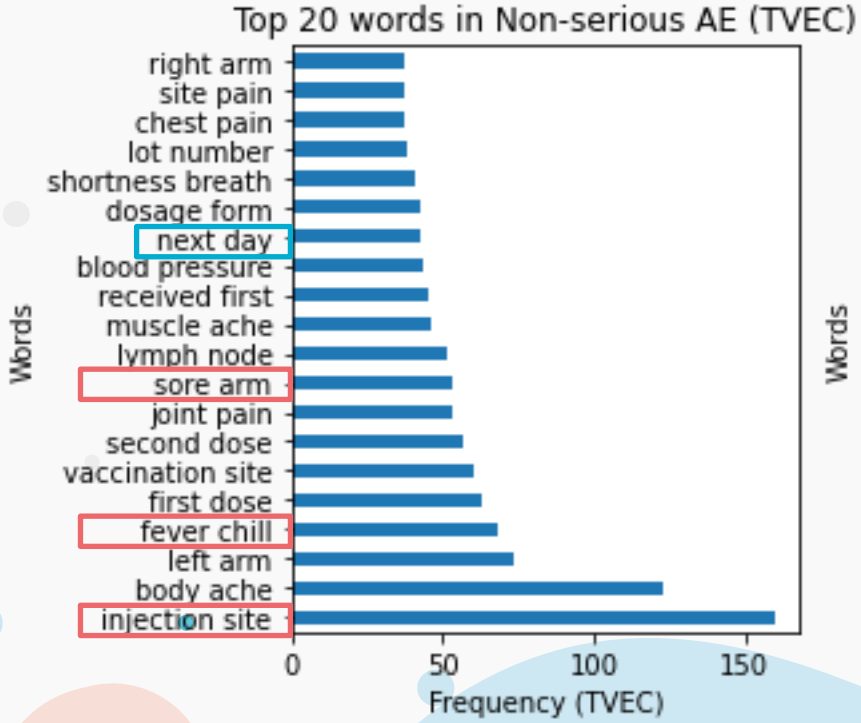
```
stopword_list = ['moderna', 'covid', 'mrna', 'vaccine', 'the', 'patient', 'pfizer',
                 'biontech', 'nan', 'none', 'mg', 'medical', 'history', 'allergy',
                 'year', 'old']
```

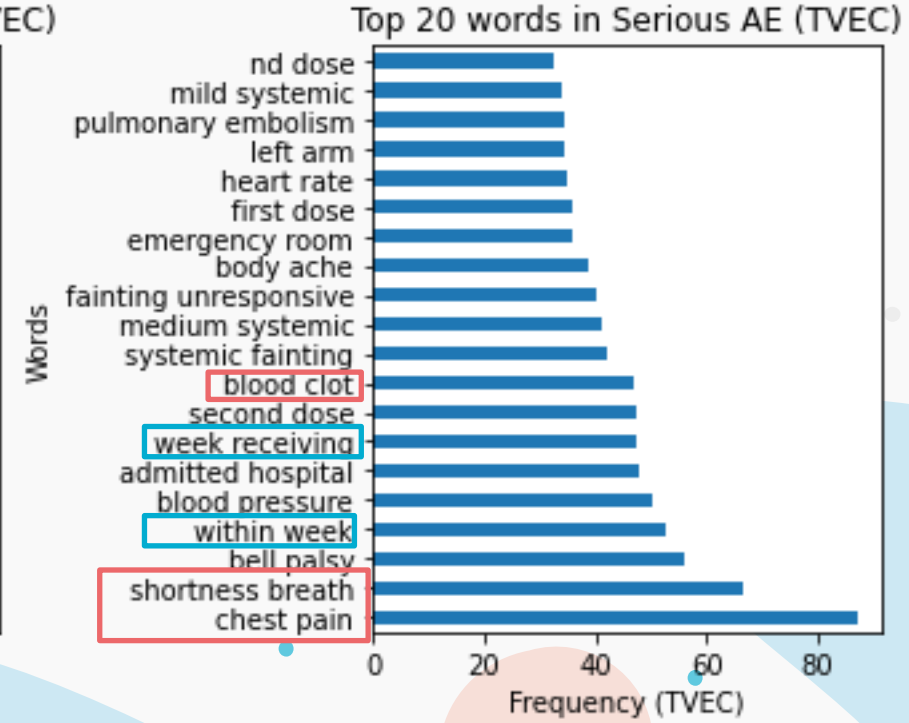
Unigram – symptom_text



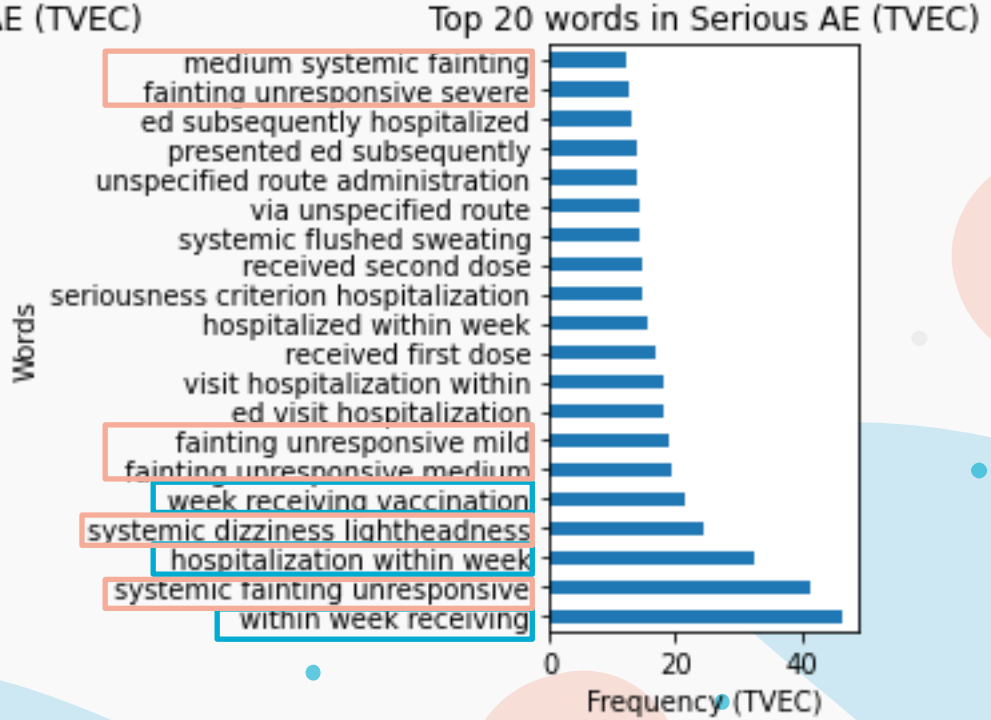


Bigram – symptom_text

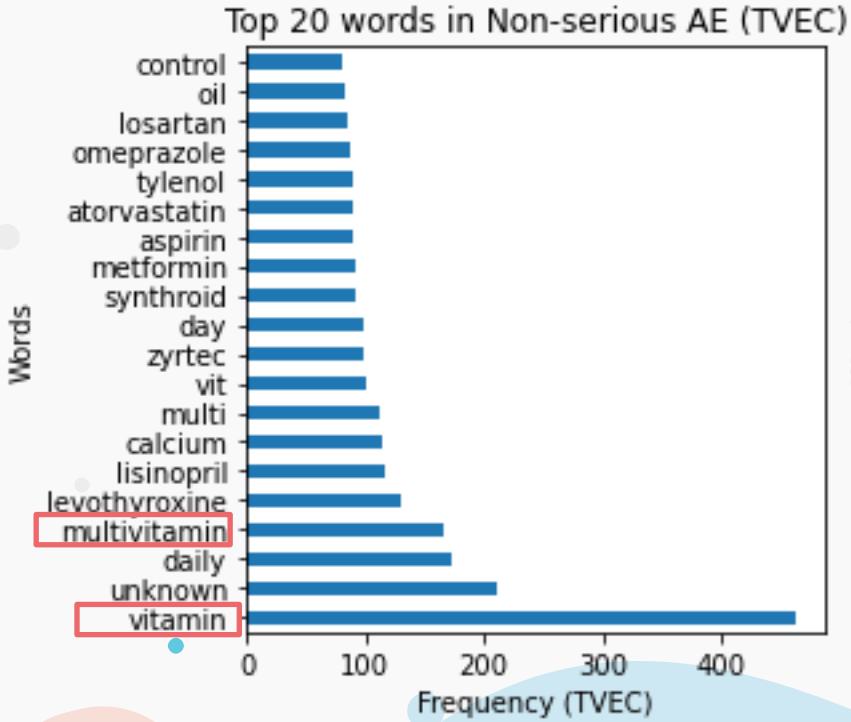


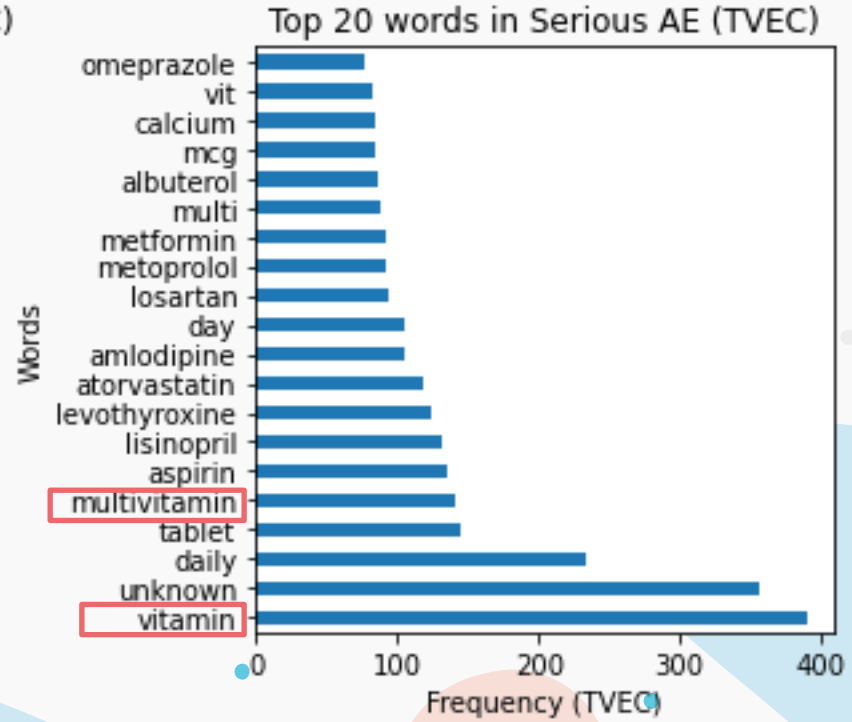


Trigram – symptom_text

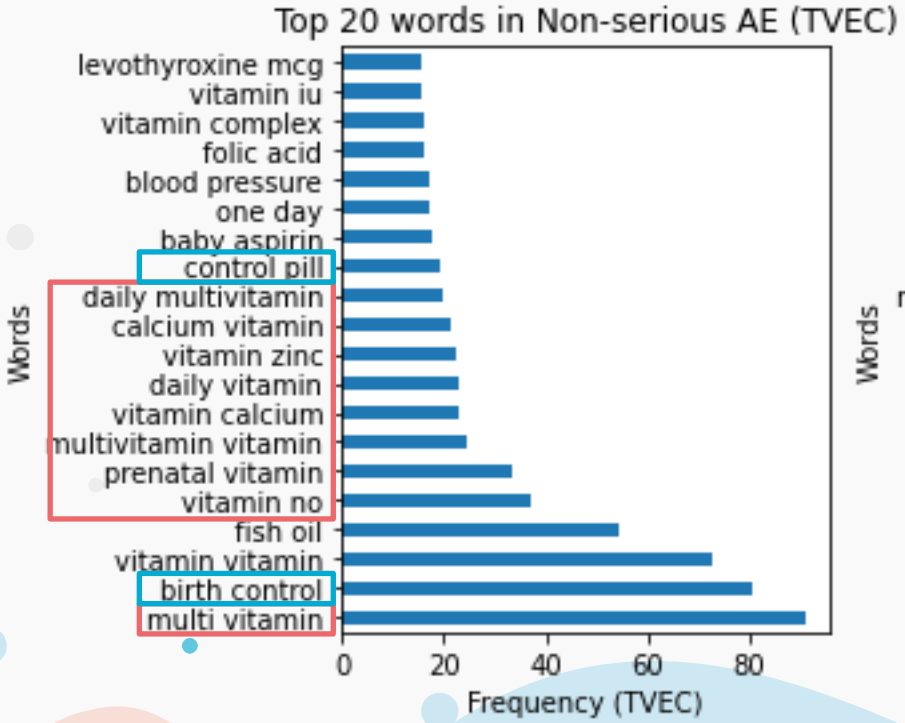


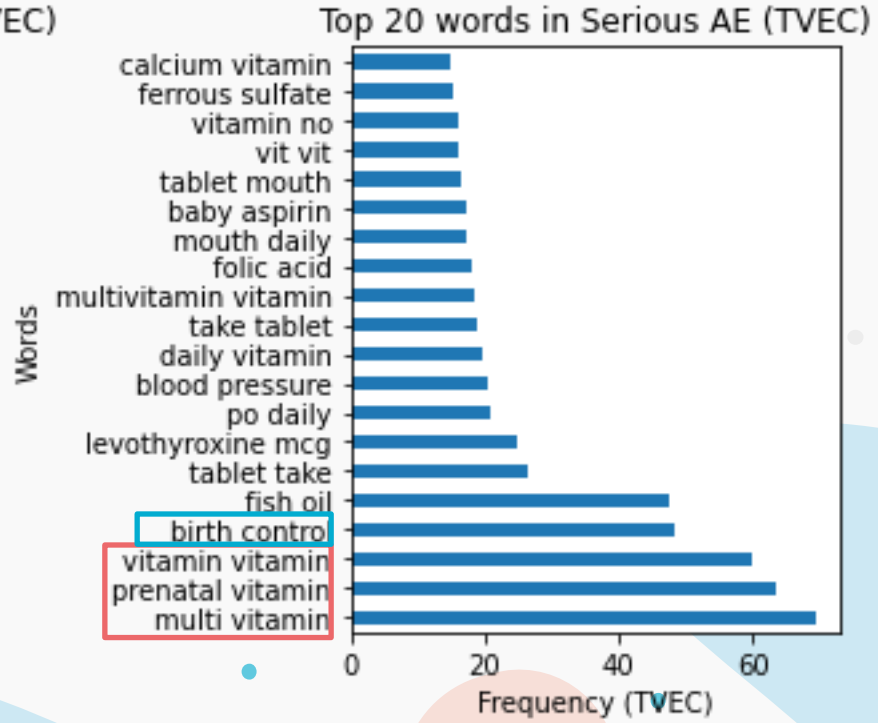
Unigram – other_meds





Bigram – other_meds

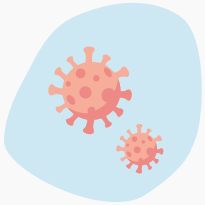




DATASET AFTER EDA

‘symptom_text’

Column for NLP



‘serious’

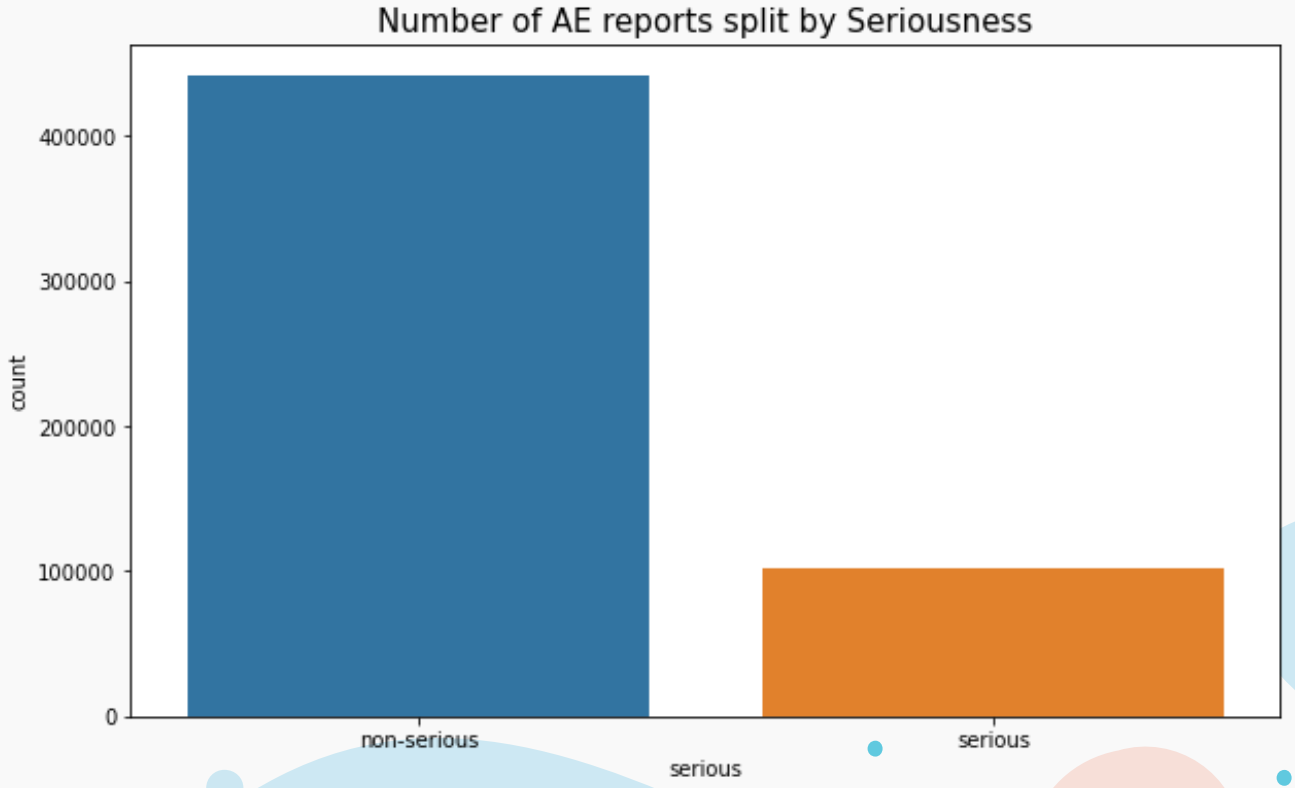
Target Column

03

MODELLING

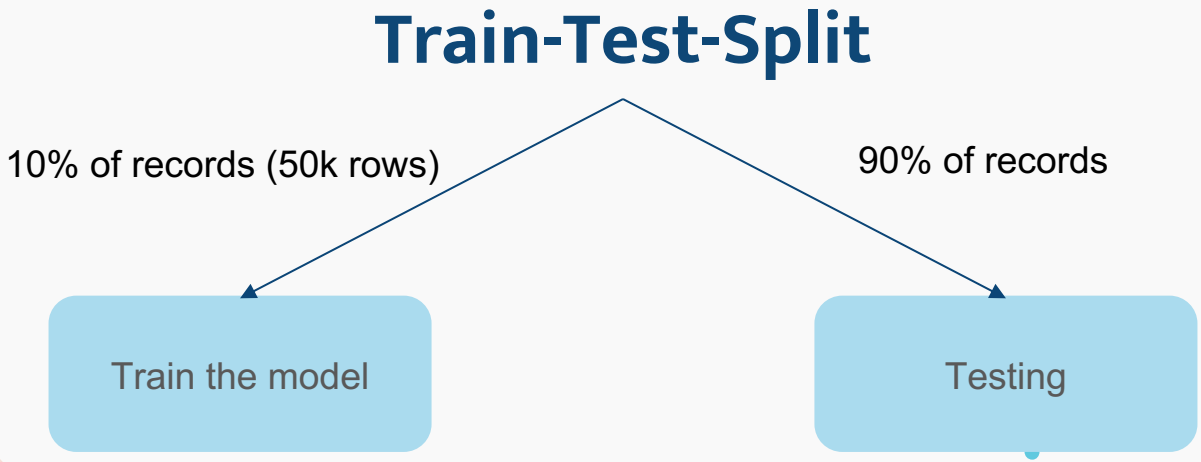
Modelling
Optimisation
Model Evaluation

Imbalanced Dataset



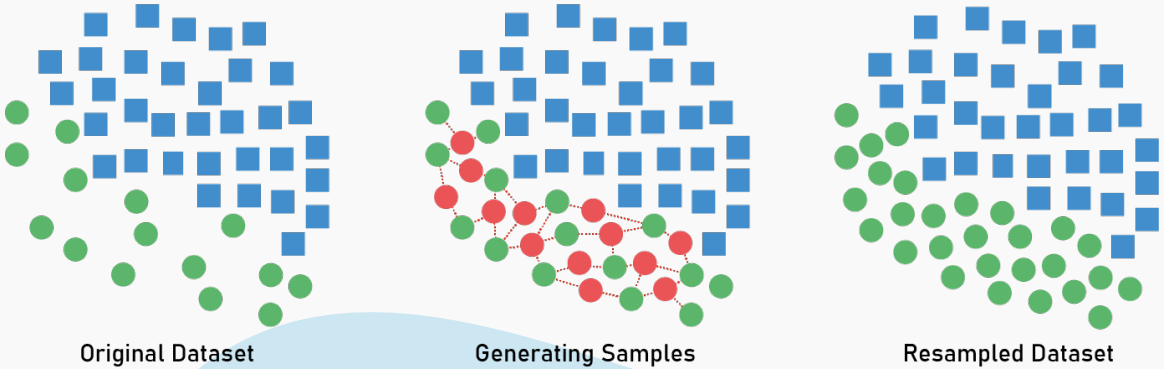
MODELLING

Hyperparameter tuning done via GridSearchCV



SMOTE

Synthetic Minority Oversampling Technique



MODELS USED

Logistic Regression (LR)

Fits data on a sigmoid curve to distinguish the 2 categories

Ada Boost Classifier (ADA)

Combines weak classifiers into a single strong classifier

Multinomial Naïve Bayes (MNB)

Uses conditional probability for classification

Support Vector Machine (SVM)

Create a hyperplane between the 2 categories

Random Forest Classifier (RFC)

Ensemble of decision trees to vote on the predicted class

MODEL RESULTS

Without SMOTE

Model No.	Word Vectorizer	Classifier	CV Score (train)	Accuracy (train)	Accuracy (test)	Recall (test)	F1 score (test)	Specificity (test)	Precision (test)
1	CountVectorizer()	LogisticRegression()	0.883	0.950	0.885	0.562	0.649	0.960	0.768
2	TfidfVectorizer()	LogisticRegression()	0.887	0.948	0.889	0.584	0.667	0.961	0.776
3	CountVectorizer()	MultinomialNB()	0.823	0.847	0.825	0.761	0.623	0.840	0.527
4	TfidfVectorizer()	MultinomialNB()	0.882	0.899	0.884	0.514	0.628	0.971	0.806
5	CountVectorizer()	RandomForestClassifier()	0.871	0.980	0.872	0.584	0.634	0.939	0.693
6	TfidfVectorizer()	RandomForestClassifier()	0.881	0.980	0.883	0.542	0.637	0.963	0.773
7	CountVectorizer()	AdaBoostClassifier()	0.861	0.861	0.860	0.333	0.473	0.983	0.819
8	TfidfVectorizer()	AdaBoostClassifier()	0.861	0.865	0.861	0.348	0.487	0.981	0.810
9	CountVectorizer()	SVC()	0.846	0.896	0.849	0.287	0.419	0.981	0.777
10	TfidfVectorizer()	SVC()	0.881	0.971	0.885	0.464	0.604	0.983	0.865

MODEL RESULTS

With SMOTE

Model No.	Word Vectorizer	Classifier	CV Score (train)	Accuracy (train)	Accuracy (test)	Recall (test)	F1 score (test)	Specificity (test)	Precision (test)
1	CountVectorizer()	LogisticRegression()	0.815	0.938	0.816	0.710	0.594	0.841	0.511
2	TfidfVectorizer()	LogisticRegression()	0.837	0.933	0.839	0.732	0.633	0.864	0.558
3	CountVectorizer()	MultinomialNB()	0.826	0.854	0.833	0.753	0.631	0.851	0.542
4	TfidfVectorizer()	MultinomialNB()	0.829	0.860	0.831	0.761	0.630	0.847	0.538
5	CountVectorizer()	RandomForestClassifier()	0.797	0.952	0.803	0.652	0.557	0.839	0.486
6	TfidfVectorizer()	RandomForestClassifier()	0.829	0.963	0.834	0.682	0.609	0.870	0.550
7	CountVectorizer()	AdaBoostClassifier()	0.616	0.624	0.621	0.803	0.446	0.578	0.308
8	TfidfVectorizer()	AdaBoostClassifier()	0.584	0.594	0.588	0.847	0.438	0.528	0.296
9	CountVectorizer()	SVC()	0.694	0.799	0.704	0.763	0.494	0.690	0.365
10	TfidfVectorizer()	SVC()	0.855	0.971	0.863	0.650	0.643	0.913	0.636

MODEL COMPARISON

Top models from SMOTE/non-SMOTE

Test (Train)	CV Score	Accuracy	Recall	F1 Score	Precision
LR (TVEC) No SMOTE	0.887	0.889 (0.948)	0.584	0.667	0.776
SVC (TVEC) SMOTE	0.855	0.863 (0.971)	0.650	0.643	0.636

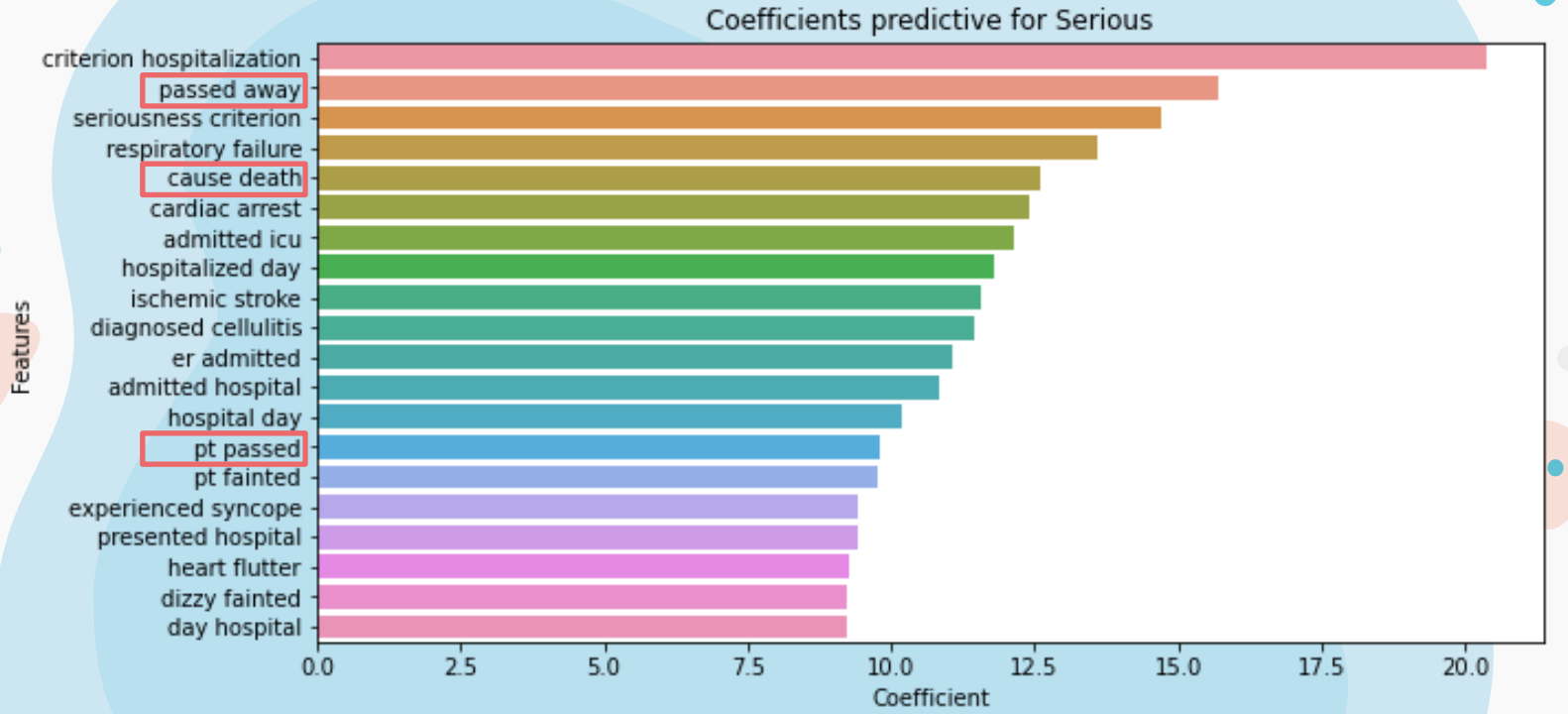
MODEL COMPARISON

Tuning of Train-Test-Split

Test (Train)	CV Score	Accuracy	Recall	F1 Score	Precision
LR (TVEC) 10/90 split	0.887	0.889 (0.948)	0.584	0.667	0.776
LR (TVEC) 80/20 split	0.902	0.903 (0.914)	0.615	0.708	0.830

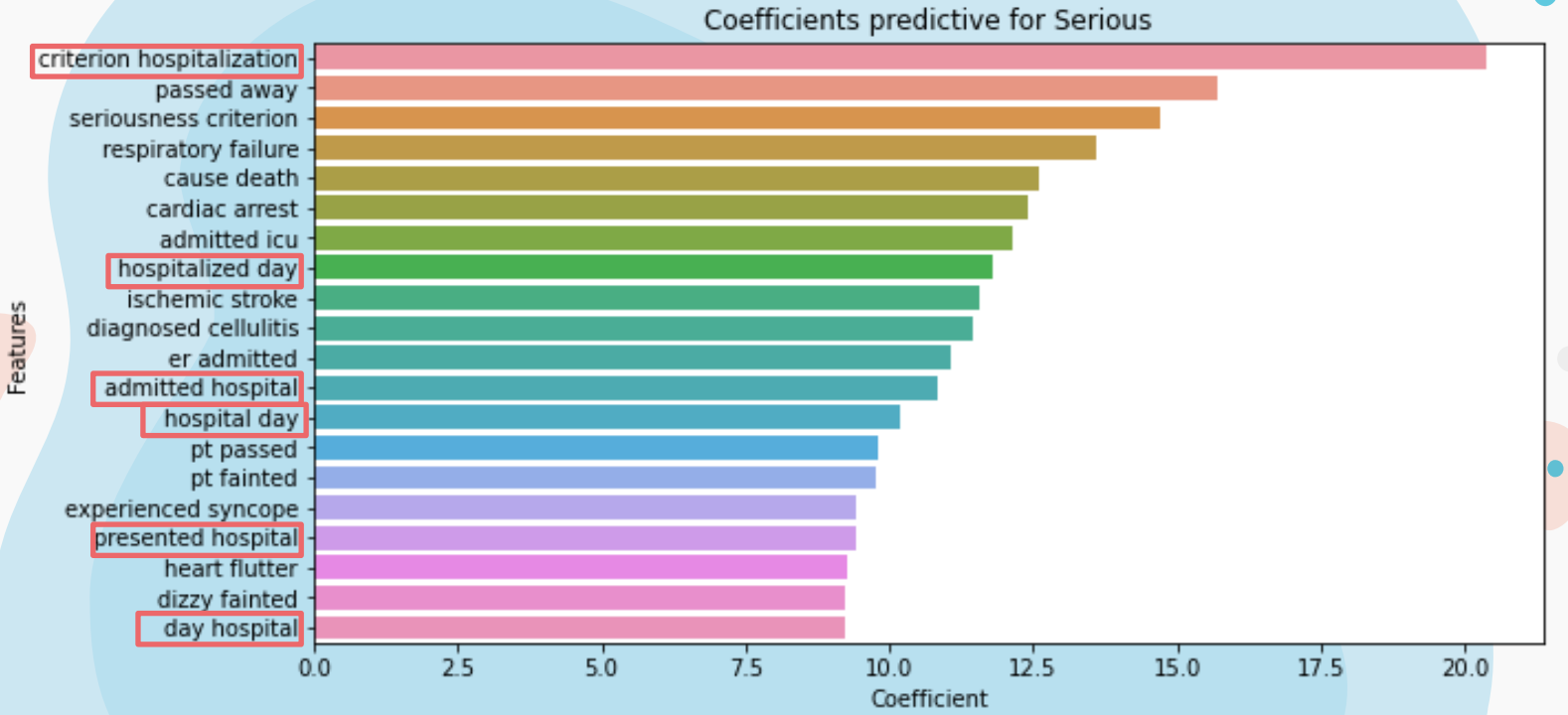
Highest Coefficients

- Death



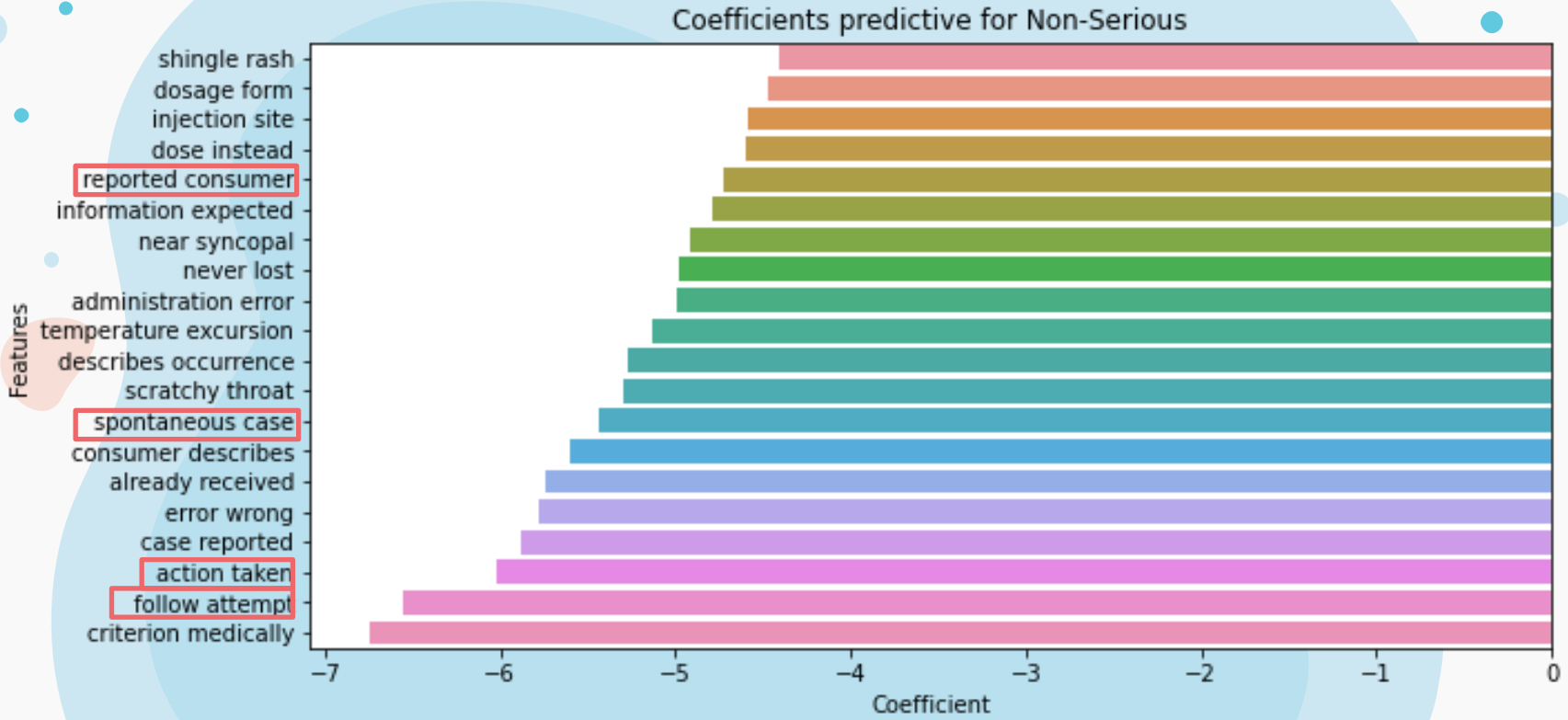
Highest Coefficients

- Hospitalisation



Lowest Coefficients

- General words that could be applied to both categories



04

CONCLUSION & RECOMMENDATION

Conclusion
Recommendation
Future Plans

CONCLUSION

Accurate Prediction
With a test and train accuracy of at least 90%, the model is able to accurately predict serious AE.

Little Overfitting
With difference between train and test accuracy at 1.1%, the model is not overfitted to the train data.

Easy to Understand
The coefficients from the model is easy to interpret and understand.

Test (Train)	CV Score	Accuracy	Recall	F1 Score	Precision
LR (TVEC) 80/20 split	90.2%	90.3% (91.4%)	61.5%	70.8%	83.0%

RECOMMENDATION

Deploy model as a preliminary screening tool for all incoming AE, allows serious cases to be labelled more quickly thus enabling signal detection to occur more efficiently

FUTURE IMPROVEMENTS

Expand Data Collection to Non-AE reports

Would allow expansion of classification scope of model (AE vs non-AE report).

Use of Deep Learning

Explore use of Neural Networks to possibly improve predictive power.

Incorporate Non-Text Columns

Use of non-text columns in model to improve model performance.

THANKS!

Do you have any questions?

facebook.com/Freepik

@Freepik_Vectors

company/freepik-company

Please keep this slide for attribution

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Stories

APPENDIX

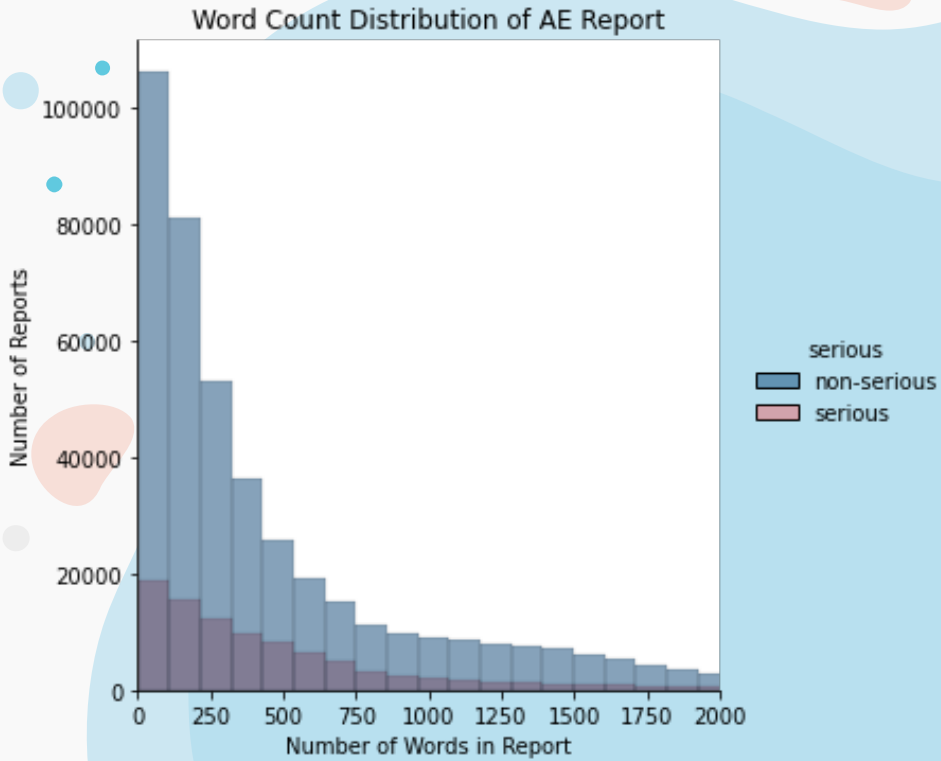
- Regex

	symptom_text	other_meds	history	allergies
17	Left side of face became numb, including to behind the left ear. Happened within 10 minutes of injection. Subsided within 30 minutes. The next day, some numbness returned at about 9pm in the evening. Pain behind left ear.	levothyroxine 100mcg/day, estradiol 1mg/day	Graves Disease	penicillin, toradol, methimazole
57	Vertigo every evening when lying down and every morning when getting up. I have been lying in bed for 5-10 minutes with eyes open, then sitting up slowly. Next, I sit on the side of the bed for a few minutes. When I get up, I need to hold onto something so I don't fall down.	multivitamin, D3, baby aspirin	none	latex, sulfa drugs
138	body aches and stomach ache	Triamterene HCTZ Montelukast Celecoxib Aller-Tec Multivitamin Vitamin D3 Magneseum	asthma when I get a cold	too much cordosone
821	12/31/2020 H/a, diarrhea, SEVERE joint pain all through body, severe exhaustion., nausea, chills, fever 99.9. It felt almost identical to my first couple says of covid.	None	Serious episode of covid + 11/18/2020	None
822	12/31/2020 H/a, diarrhea, SEVERE joint pain all through body, severe exhaustion., nausea, chills, fever 99.9. It felt almost identical to my first couple says of covid.	None	Serious episode of covid + 11/18/2020	None

APPENDIX

- Stemming vs Lemmatizing

	symptom_text_stemmed	symptom_text_lemmatized
0	[left, side, of, face, becam, numb, includ, to...	[Left, side, of, face, became, numb, including...
1	[vertigo, everi, even, when, lie, down, and, e...	[Vertigo, every, evening, when, lying, down, a...
2	[bodi, ach, and, stomach, ach]	[body, ache, and, stomach, ache]
3	[, h, a, diarrhea, sever, joint, pain, all, th...	[, H, a, diarrhea, SEVERE, joint, pain, all, t...
4	[, h, a, diarrhea, sever, joint, pain, all, th...	[, H, a, diarrhea, SEVERE, joint, pain, all, t...
5	[moderna, covid, vaccin, eua, headach, nausea,...	[Moderna, covid, vaccine, EUA, headache, nause...
6	[moderna, covid, vaccin, eua, headach, nausea,...	[Moderna, covid, vaccine, EUA, headache, nause...
7	[no, advers, reaction, the, staff, were, given...	[No, adverse, reaction, the, staff, were, give...
8	[moderna, covid, vaccin, eua, flare, up, of, g...	[Moderna, COVID, Vaccine, EUA, Flare, up, of, ...
9	[moderna, covid, vaccin, eua, flare, up, of, g...	[Moderna, COVID, Vaccine, EUA, Flare, up, of, ...
10	[moderna, covid, vaccin, eua, fever, of, for, ...	[Moderna, COVID, Vaccine, EUA, Fever, of, for,...

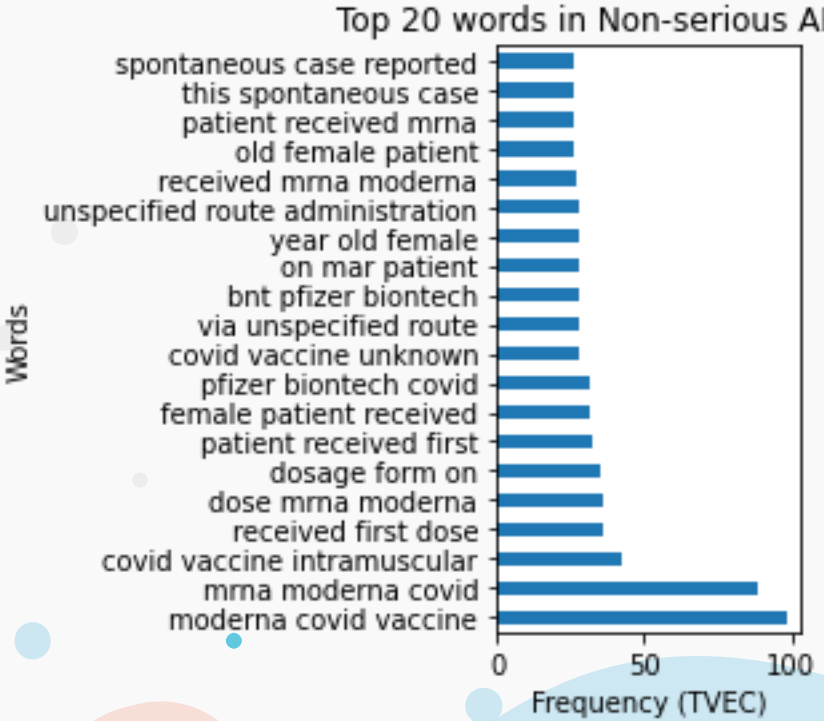


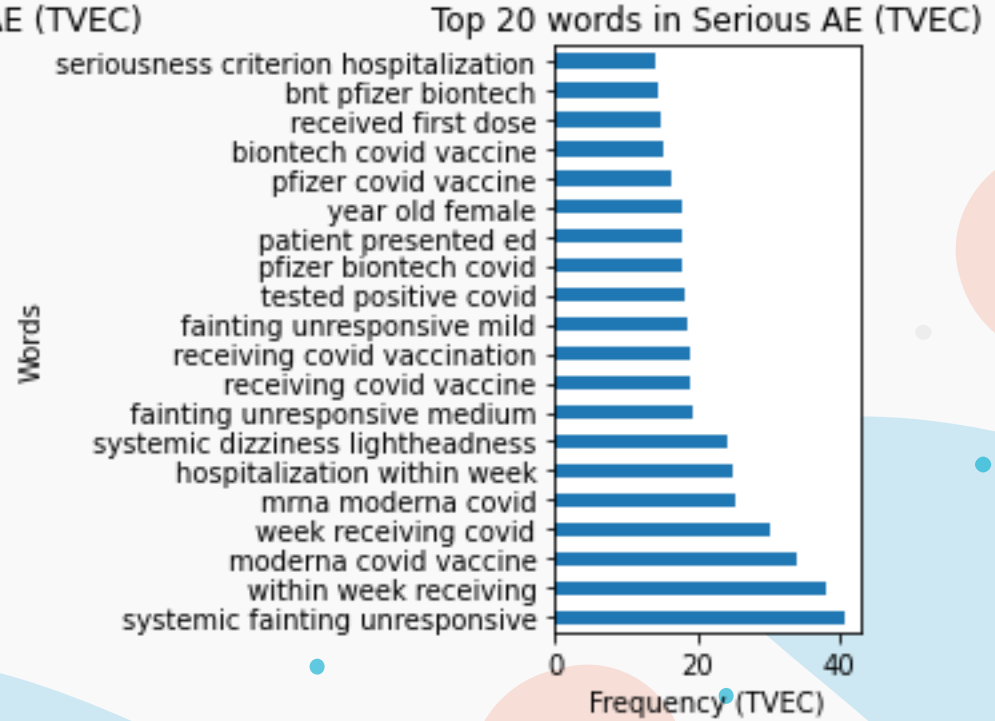
Word Vectorization

- Count Vectorization
- TF-IDF Vectorization

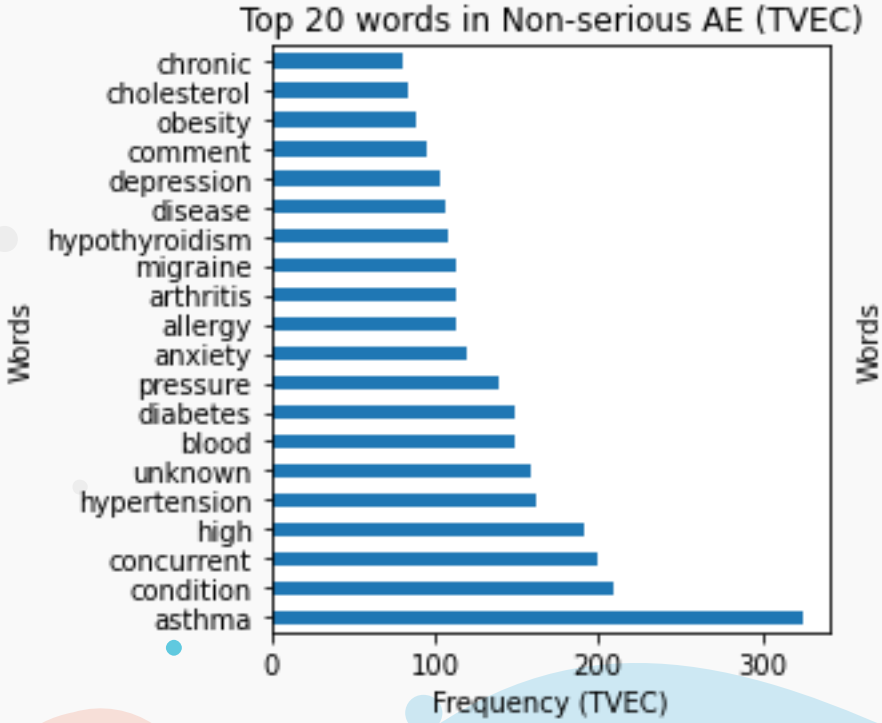
- N-gram frequency

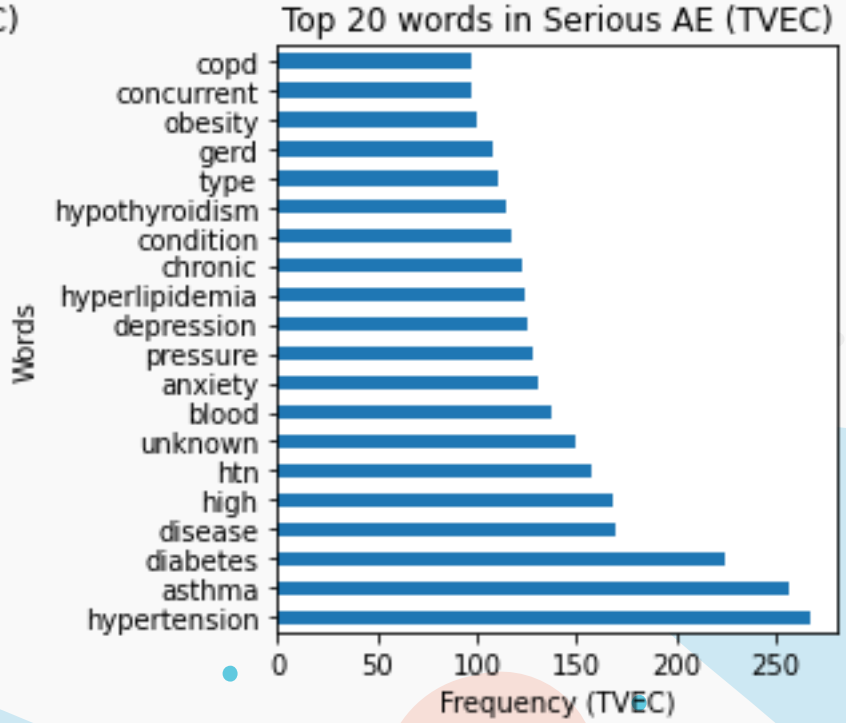
Trigram – symptom_text



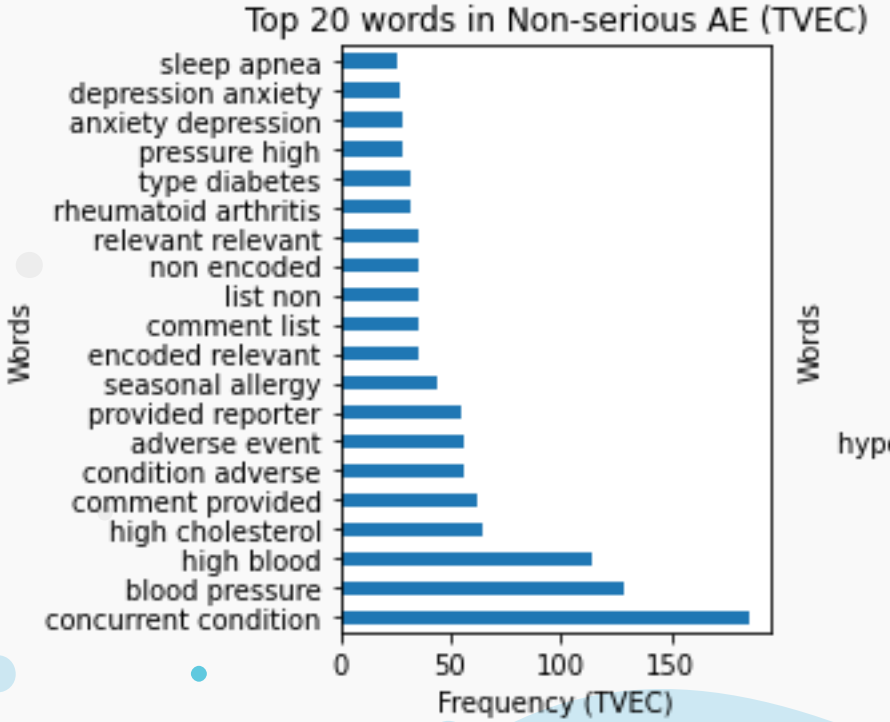


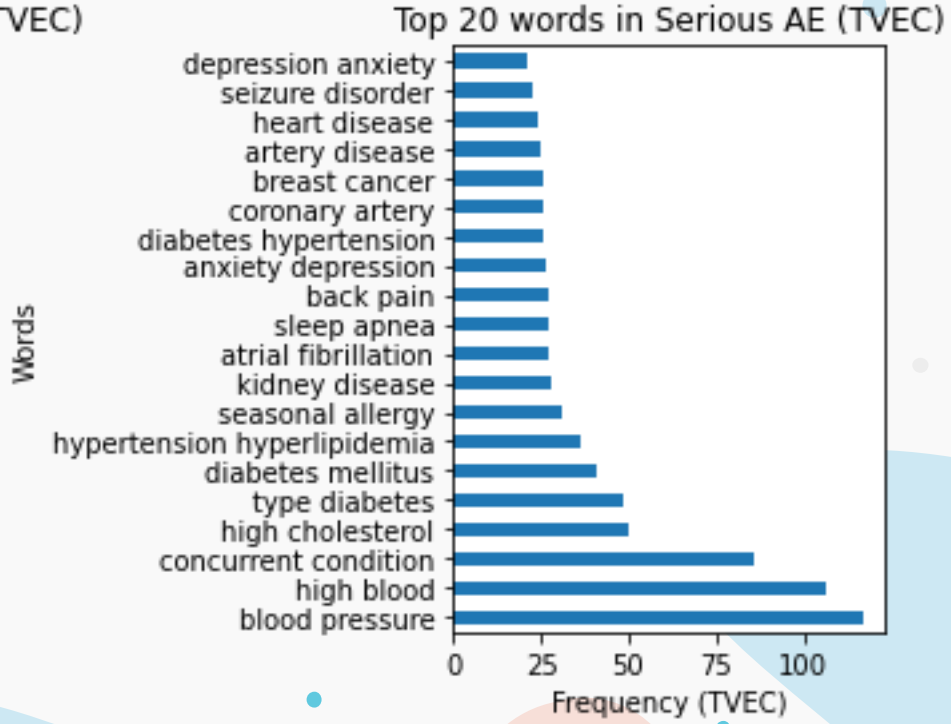
Unigram – history



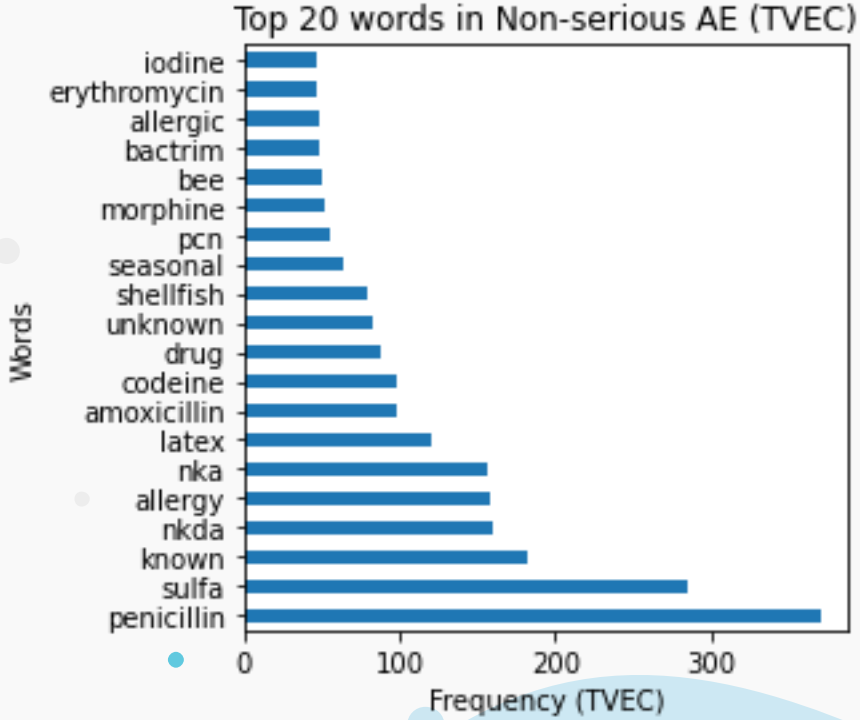


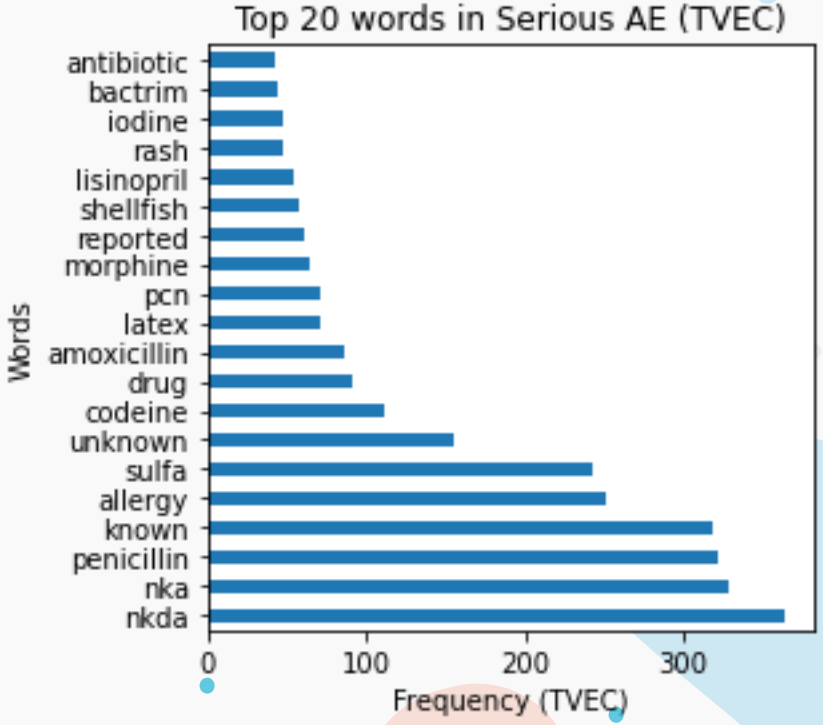
Bigram – history



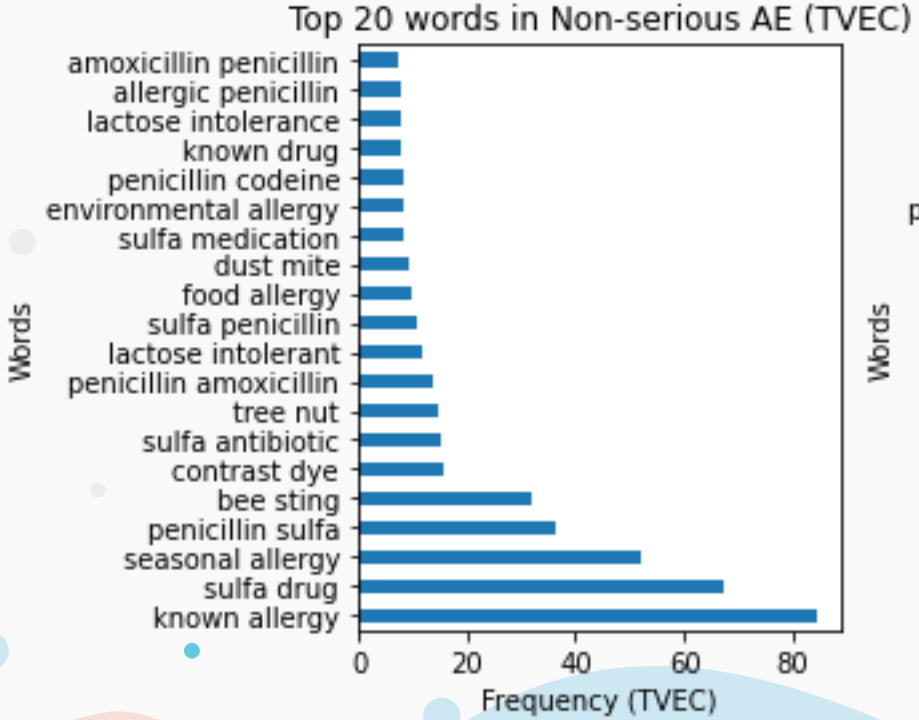


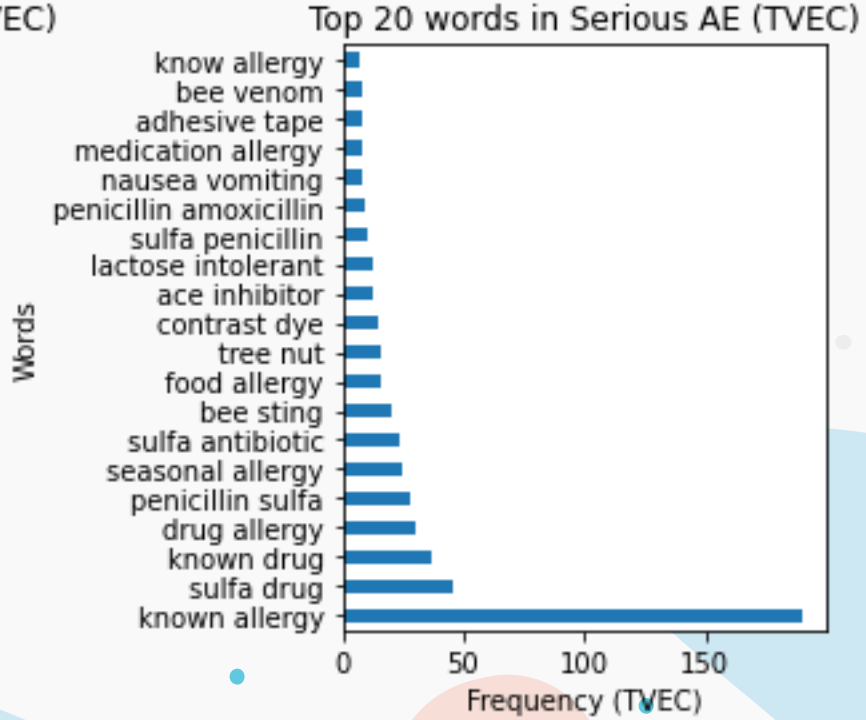
Unigram – allergies





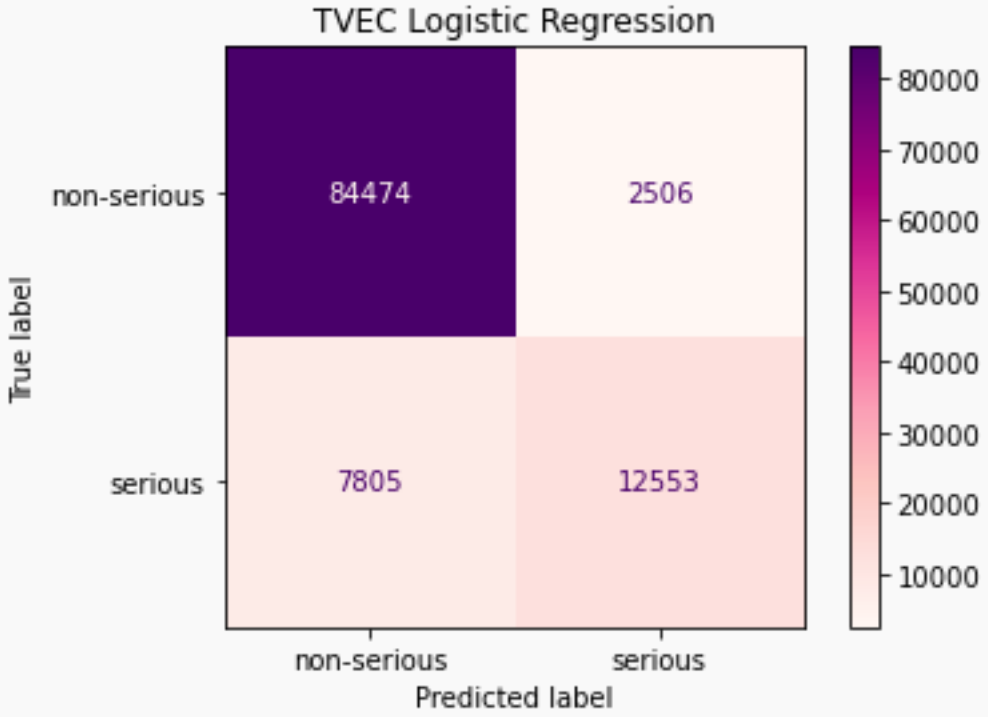
Bigram – allergies





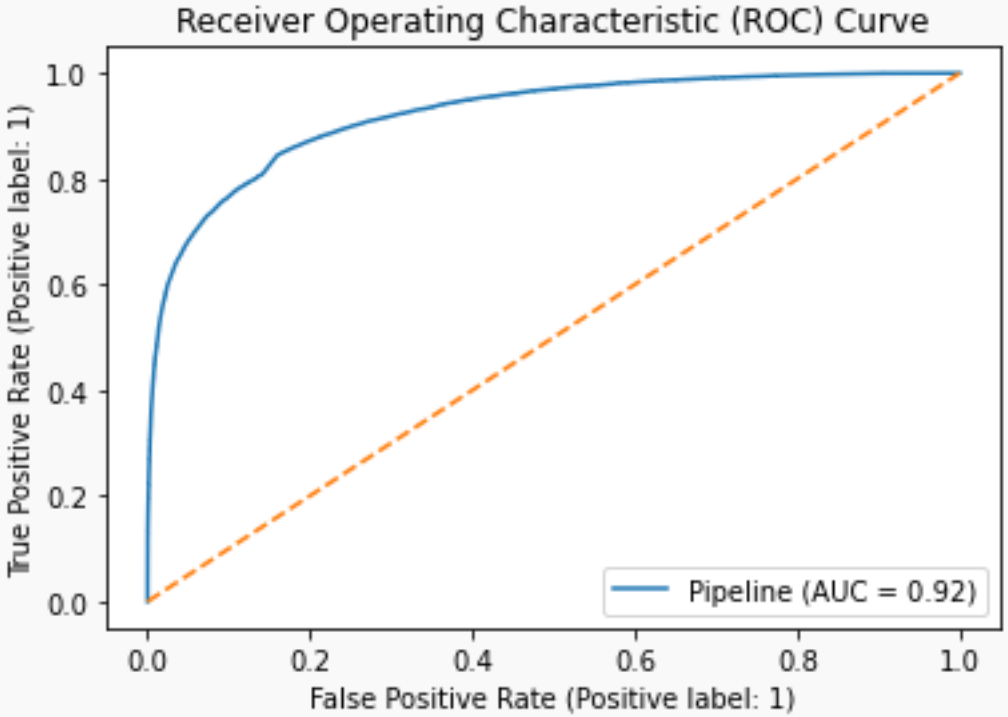
APPENDIX

- Confusion Matrix



APPENDIX

- ROC-AUC



APPENDIX

All models

Model No.	SMOTE?	Word Vectorizer	Classifier	CV Score (train)	Accuracy (train)	Accuracy (test)	Recall (test)	F1 score (test)	Specificity (test)	Precision (test)
1	No	CountVectorizer	LogisticRegression	0.883	0.95	0.885	0.562	0.649	0.96	0.768
2	No	TfidfVectorizer	LogisticRegression	0.887	0.948	0.889	0.584	0.667	0.961	0.776
3	No	CountVectorizer	MultinomialNB	0.823	0.847	0.825	0.761	0.623	0.84	0.527
4	No	TfidfVectorizer	MultinomialNB	0.882	0.899	0.884	0.514	0.628	0.971	0.806
5	No	CountVectorizer	RandomForestClassifier	0.871	0.98	0.872	0.584	0.634	0.939	0.693
6	No	TfidfVectorizer	RandomForestClassifier	0.881	0.98	0.883	0.542	0.637	0.963	0.773
7	No	CountVectorizer	AdaBoostClassifier	0.861	0.861	0.86	0.333	0.473	0.983	0.819
8	No	TfidfVectorizer	AdaBoostClassifier	0.861	0.865	0.861	0.348	0.487	0.981	0.81
19	No	CountVectorizer	SVC	0.846	0.896	0.849	0.287	0.419	0.981	0.777
10	No	TfidfVectorizer	SVC	0.881	0.971	0.885	0.464	0.604	0.983	0.865
11	Yes	CountVectorizer	LogisticRegression	0.815	0.938	0.816	0.71	0.594	0.841	0.511
12	Yes	TfidfVectorizer	LogisticRegression	0.837	0.933	0.839	0.732	0.633	0.864	0.558
13	Yes	CountVectorizer	MultinomialNB	0.826	0.854	0.833	0.753	0.631	0.851	0.542
14	Yes	TfidfVectorizer	MultinomialNB	0.829	0.86	0.831	0.761	0.63	0.847	0.538
15	Yes	CountVectorizer	RandomForestClassifier	0.797	0.952	0.803	0.652	0.557	0.839	0.486
16	Yes	TfidfVectorizer	RandomForestClassifier	0.829	0.963	0.834	0.682	0.609	0.87	0.55
17	Yes	CountVectorizer	AdaBoostClassifier	0.616	0.624	0.621	0.803	0.446	0.578	0.308
18	Yes	TfidfVectorizer	AdaBoostClassifier	0.584	0.594	0.588	0.847	0.438	0.528	0.296
19	Yes	CountVectorizer	SVC	0.694	0.799	0.704	0.763	0.494	0.69	0.365
20	Yes	TfidfVectorizer	SVC	0.855	0.971	0.863	0.65	0.643	0.913	0.636
21	No	CountVectorizer	LogisticRegression (larger train dataset)	0.898	0.913	0.898	0.586	0.685	0.971	0.824
22	No	TfidfVectorizer	LogisticRegression (larger train dataset)	0.902	0.914	0.903	0.617	0.708	0.971	0.83
23	No	CountVectorizer	MultinomialNB (larger train dataset)	0.827	0.83	0.827	0.782	0.631	0.837	0.53
24	No	TfidfVectorizer	MultinomialNB (larger train dataset)	0.887	0.89	0.887	0.633	0.679	0.946	0.733

APPENDIX

- 80/20 train-test split

Model No.	Word Vectorizer	Classifier	CV Score (train)	Accuracy (train)	Accuracy (test)	Recall (test)	F1 score (test)	Specificity (test)	Precision (test)
1	CountVectorizer()	LogisticRegression()	0.898	0.913	0.898	0.586	0.685	0.971	0.824
2	TfidfVectorizer()	LogisticRegression()	0.902	0.914	0.903	0.617	0.708	0.971	0.830
3	CountVectorizer()	MultinomialNB()	0.827	0.830	0.827	0.782	0.631	0.837	0.530
4	TfidfVectorizer()	MultinomialNB()	0.887	0.890	0.887	0.633	0.679	0.946	0.733