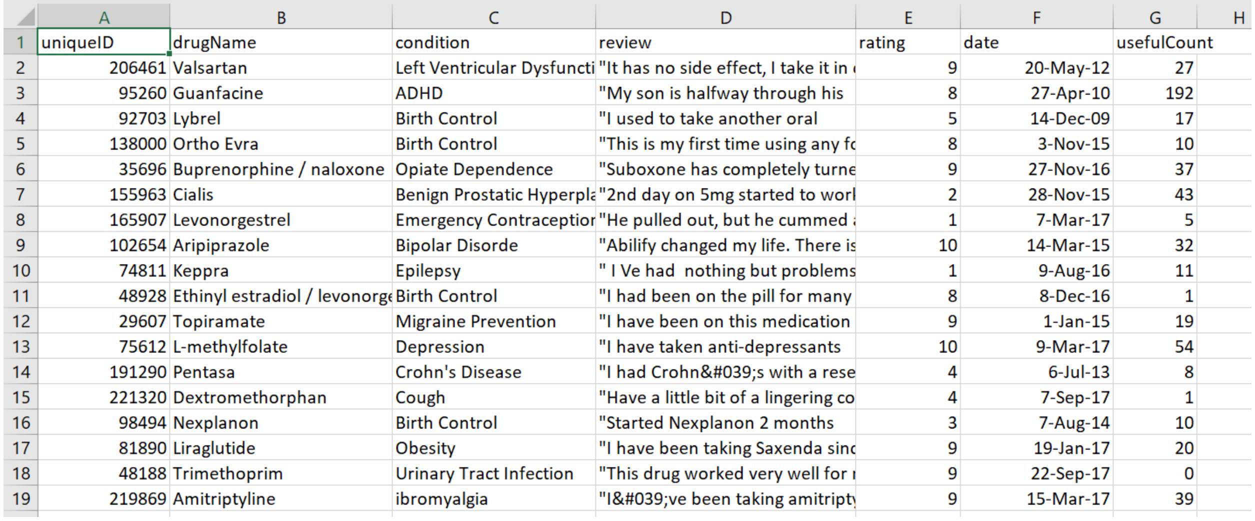
SCENARIO-3

SENTIMENT ANALYSIS- CUSTOM MODEL

Sentiment Analysis is the NLP technique that performs on the text to determine whether the author's intentions towards a particular topic, product, etc. are positive, negative, or neutral.

Sentiment analysis can help us attain the attitude and mood of the wider public which can then help us gather insightful information about the context.

- Here we have used 30k dataset.
- We have taken our dataset from kaggle. Link- [UCI ML Drug Review dataset | Kaggle]
- Our dataset is as follows:



	A	B	C	D	E	F	G	H
1	uniqueID	drugName	condition	review	rating	date	usefulCount	
2	206461	Valsartan	Left Ventricular Dysfuncti	"It has no side effect, I take it in	9	20-May-12	27	
3	95260	Guanfacine	ADHD	"My son is halfway through his	8	27-Apr-10	192	
4	92703	Lybrel	Birth Control	"I used to take another oral	5	14-Dec-09	17	
5	138000	Ortho Evra	Birth Control	"This is my first time using any fo	8	3-Nov-15	10	
6	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turne	9	27-Nov-16	37	
7	155963	Cialis	Benign Prostatic Hyperpla	"2nd day on 5mg started to worl	2	28-Nov-15	43	
8	165907	Levonorgestrel	Emergency Contraceptior	"He pulled out, but he cummed	1	7-Mar-17	5	
9	102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life. There is	10	14-Mar-15	32	
10	74811	Keppra	Epilepsy	" I Ve had nothing but problems	1	9-Aug-16	11	
11	48928	Ethinyl estradiol / levonorge	Birth Control	"I had been on the pill for many	8	8-Dec-16	1	
12	29607	Topiramate	Migraine Prevention	"I have been on this medication	9	1-Jan-15	19	
13	75612	L-methylfolate	Depression	"I have taken anti-depressants	10	9-Mar-17	54	
14	191290	Pentasa	Crohn's Disease	"I had Crohn&#039;s with a rese	4	6-Jul-13	8	
15	221320	Dextromethorphan	Cough	"Have a little bit of a lingering co	4	7-Sep-17	1	
16	98494	Nexplanon	Birth Control	"Started Nexplanon 2 months	3	7-Aug-14	10	
17	81890	Liraglutide	Obesity	"I have been taking Saxenda sinc	9	19-Jan-17	20	
18	48188	Trimethoprim	Urinary Tract Infection	"This drug worked very well for	9	22-Sep-17	0	
19	219869	Amitriptyline	ibromyalgia	"I&#039;ve been taking amitripty	9	15-Mar-17	39	

STEPS:

1. This code is written in python so we have to import the following libraries.
   - import pandas as pd
   - import numpy as np

    - from nltk.stem import WordNetLemmatizer
    - import nltk
    - from nltk.corpus import stopwords
    - from nltk.tokenize import word_tokenize
    - import re
    - from nltk.corpus import sentiwordnet as swn
    - from IPython.display import clear_output

2. First, we are going to need **NumPy** and **pandas** “***Re”*** stands for **regular expression** which is used to extract a certain portion of a string. ***Nltk*** is an **NLP library** and we are going to import it in certain parts of our code to process the textual data. Then we are going to import ***sklearn*** for model creation. We are also importing some metrics from sklearn to analyze model performance.
3. With the help of **pandas** read the raw data which is provided in the form of xls/csv.
4. Here we have the **drug dataset** which we have downloaded from **kaggle**.
5. Extract the reviews column as we have to give the **sentiment** of the reviews.
6. Here we have selected the top 30k rows.
7. Now we have to **preprocess** the data [data cleaning].
    - Remove the stopwords such as a,an,the.
    - Convert all the words into lower case.
    - Join single space to all the single character words such as ‘I’.
    - Substituted the multiple spaces with single spaces
    - Removed all the single characters in the text
    - Removed the @ handlers.

8. Tokenised all the words by using **word_tokenize** function.**Tokenization** is the first step in any NLP pipeline.A tokenizer breaks unstructured data and natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document. Tokenization can separate sentences, words, characters, or subwords.
9. **Lemmatised** all the words by using **WordNetLemmatizer()**.Lemmatization aims to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.
10. Now we have the final column after the preprocessing of the data.
11. **WordNet** is a lexical database of semantic relations between words in more than 200 languages. WordNet links words into semantic relations including synonyms, hyponyms, and meronyms. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.
12. **SentiWordNet** is a lexical resource in which each WordNet synset is associated to three numerical scores Obj(s), Pos(s) and Neg(s), describing how objective, positive, and negative the terms contained in the synset are.
13. We have calculated the parts of speech tag by traversing all the rows of the reviews column .**POS Tagging (Parts of Speech Tagging)** is a process to mark up the words in text format for a particular part of a speech based on its definition and context.This means labeling words in a sentence as nouns, adjectives, verbs...etc. Even more impressive, it also labels by tense, and more.Here we have used nltk library.

14. Then we convert all the parts of speech tag into ADJ,NOUN,ADV,VERB as we have to pass this to **synset** function in order to get the scores and only these categories are of use to us.
15. **Synset** is a special kind of a simple interface that is present in NLTK to look up words in WordNet. Synset instances are the groupings of synonymous words that express the same concept. Some of the words have only one Synset and some have several.
16. We will get a synset of words and we will take the first sense.
17. We will pass this word to the **senti_synset** function to get the scores.From that we will get the positive score,negative and the neutral score of the synset that we passed.
18. Now since we got the sentiment scores for all the synsets and words we will now calculate the total sentiment score of the sentence by subtracting the negative scores from the positive scores.
19. Later we put the sentiment scores of all the sentences under the column **'senti_score'.**
20. Now we will label the sentiment as **positive,negative** or **neutral** on the basis of a threshold value.Here after comparing various threshold values we took **0.05** as the threshold.
21. We have taken score>0.05 to be positive,score<-0.05 to be negative and rest as neutral.
22. So now we have our final review column as well as the sentiment of that particular review.
23. To make sense of this data for our machine learning algorithm, we will need to convert each review to a numerical representation that we call **vectorization**.

24. Here we tried using **count vectorizer** and tf-idf vectorizer and for our dataset we found out that count-vectorizer gave better accuracy.
    Count Vectorizer is a way to convert a given set of strings into a frequency representation.
25. After vectorizing,we train_test_split our data in 20/80 ratio.
26. After splitting and vectorize text reviews into numbers, we will generate a Logistic Regression model on the training set and perform prediction on test set features.
27. Checked the correctness of the model after it has been created by comparing real and anticipated values. This model is 84 % accurate.

COMPARISON OF VARIOUS ALGORITHMS

The various Machine Learning algorithms that are commonly used for sentiment analysis are as follows:

- Support Vector Machine(SVM)
- Logistic Regression
- Naive Bayes
- Long Short Term Memory(LSTM)
- Random Forest

But we decided to compare only SVM, Naive Bayes and Logistic Regression because of the following advantages:

ADVANTAGES

Support Vector Machine(SVM):

- SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin hyper plane. Finding a hyperplane can be useful to classify the data correctly between different groups.
- SVM generally do not suffer condition of overfitting and performs well when there is a clear indication of separation between classes.
- Outliers have less influence in SVM Algorithm therefore there are less chances of skewing the results as outliers affect the mean of the data and therefore mean cannot represent the data set which it was able to do before the effect of having outliers ,Thus as there is less influence of outliers in SVM ,it proves to be helpful.
- SVMs can be robust, even when the training sample has some bias.
- Works best on small sample sets because of its high training time.
- At the outset, Support Vector Machine (SVM) classification algorithm gives almost same accuracy [83%] as logistic but takes more time.

Naive Bayes

- This algorithm works quickly and can save a lot of time.
- Naive Bayes is suitable for solving multi-class prediction problems.
- If its assumption of the independence of features holds true, it can perform better than other models and requires much less training data.

Logistic Regression :

- Logistic regression is easier to implement, interpret, and very efficient to train.
- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
- It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).
- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
- Logistic regression is less inclined to over-fitting.

DISADVANTAGES

Support Vector Machine(SVM):

- It takes a long training time when working with large datasets.
- It is hard to understand the final model and individual impact.

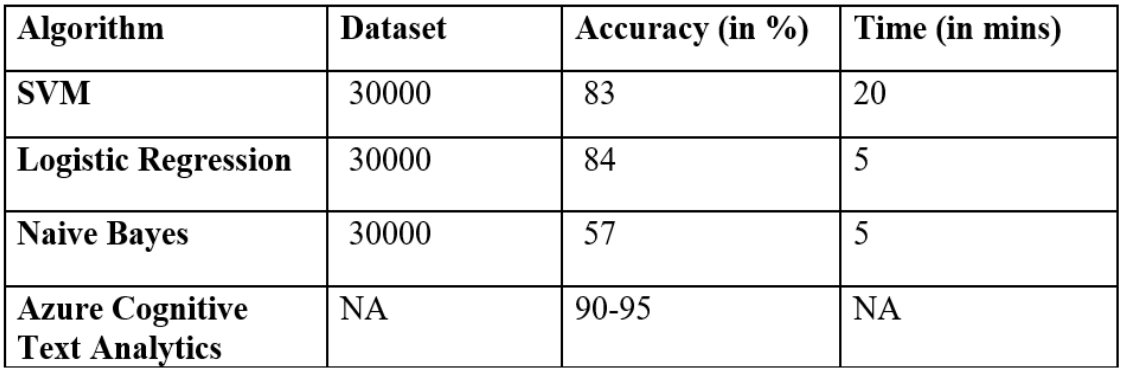
Naives Bayes:

1. Naive Bayes is that it relies on the distribution of the training dataset.
    - A good data set will contain the same proportion of positive and negative documents as a random sample would. However, most of the available in the real world are artificially balanced.

2. Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.
3. If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency.
4. Naive Bayes is vulnerable to numerical underflow in the prediction step if the dimensionality of the predictors is much larger than the number of observations.

After comparing all the algorithms on the basis of their **time** and **accuracy** we chose to go ahead with logistic regression for training out custom model.

COMPARISON TABLE



Algorithm	Dataset	Accuracy (in %)	Time (in mins)
SVM	30000	83	20
Logistic Regression	30000	84	5
Naive Bayes	30000	57	5
Azure Cognitive Text Analytics	NA	90-95	NA