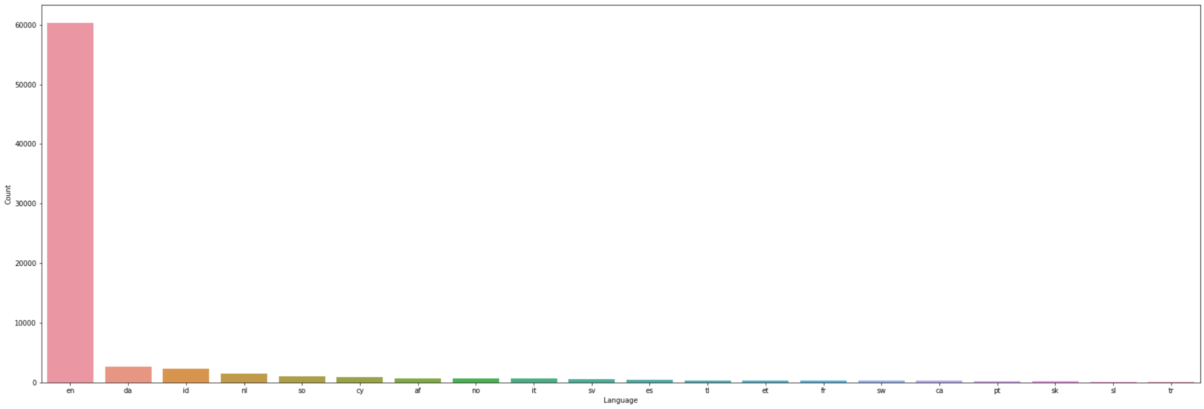
Introduction

The problem statement is about detection of help seeking tweets related to **oxygen related covid tweets**. We have been given about 70K tweets of which only about 300 tweets are labelled.So using this data we have to make a classifier that can find the tweets which are help seeking or not or are neutral. The tweets are related to COVID times.

Data

I have found the tweets contain emojis and local language like hindi , bangali or other. Emojis can signify whether people need help or not but extraction of that information is a little difficult and also will not help that much so to tell what is the expression of the people,and it is a little bit difficult to extract information from them , so for all this reason we have cleaned them up through tweets in preprocess. In cleaning we removed all emojis , and all special signs etc and lowercase the tweets

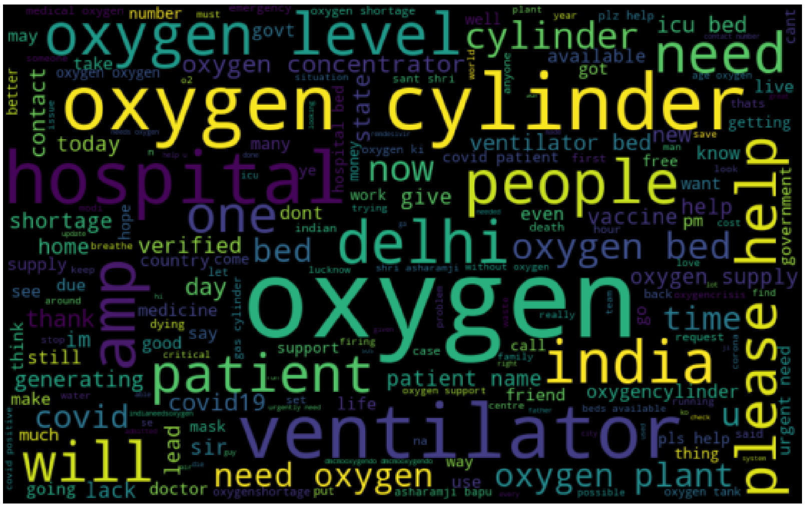
Major Language in tweets.



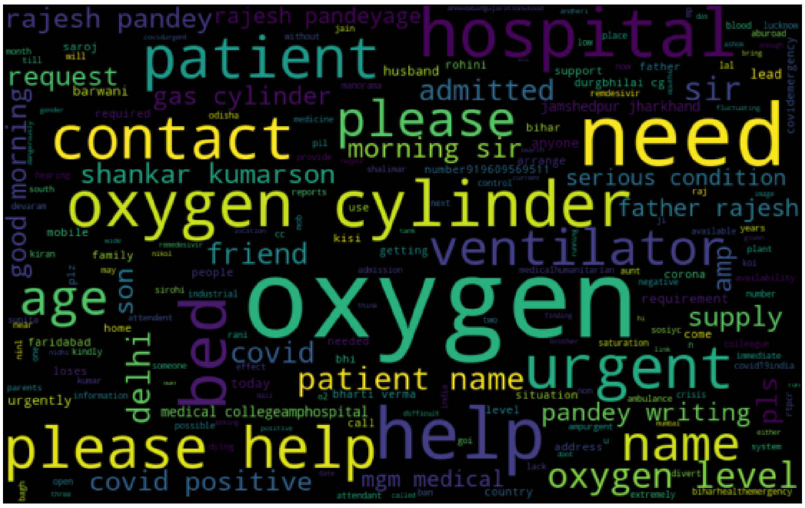
So, as you can see most of the tweets (more than 85%) are of the **English** language. So we can ignore other language tweets as it will become more complicated

Most Used Words

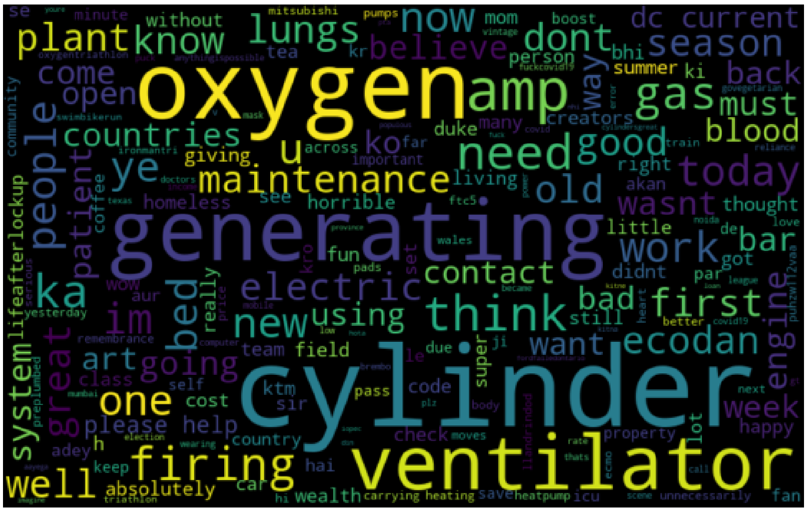
- Below is the top words used in **all tweets**



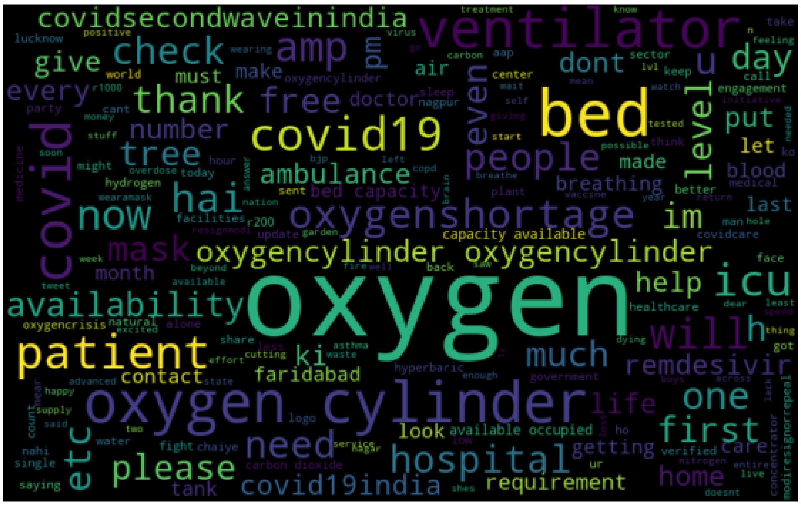
- Most words used in **help seeking tweets**



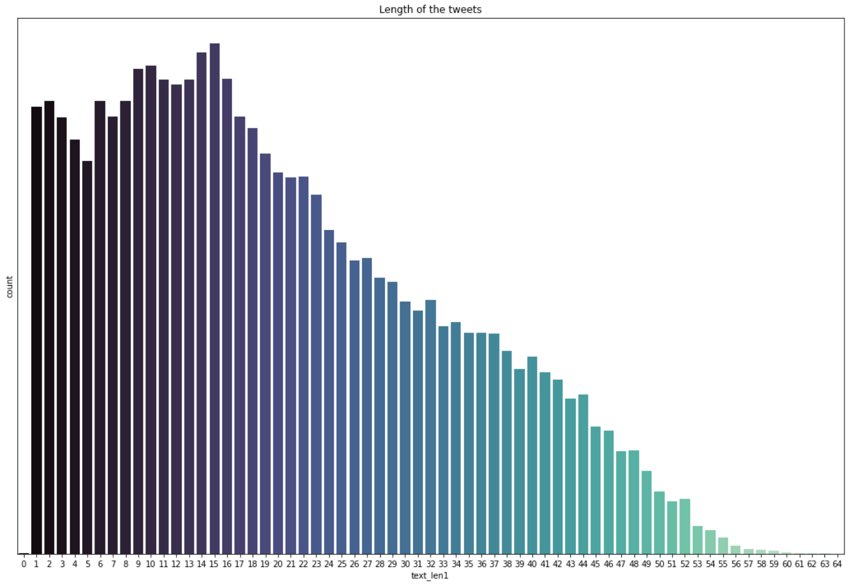
- Most words used in **tweets which are not for help**



- Most words which used in **neutral tweets**

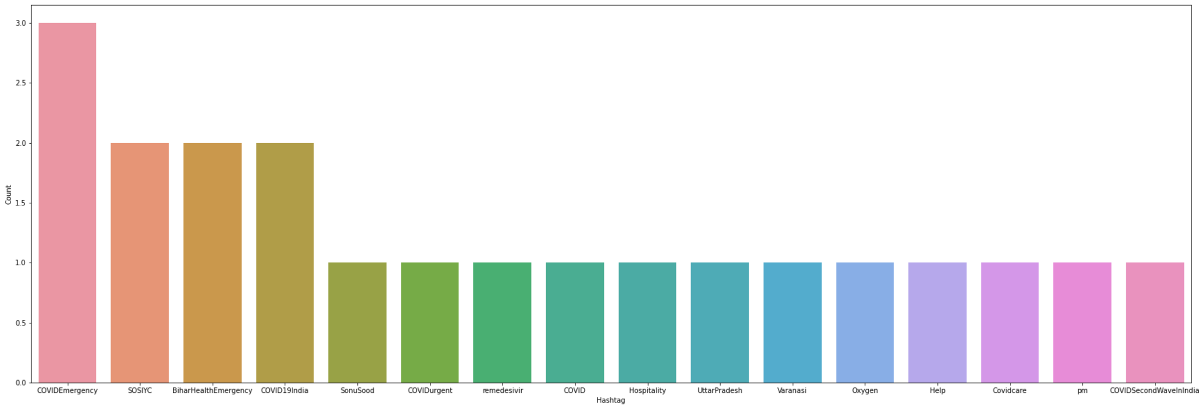


Average length of the tweets = 15

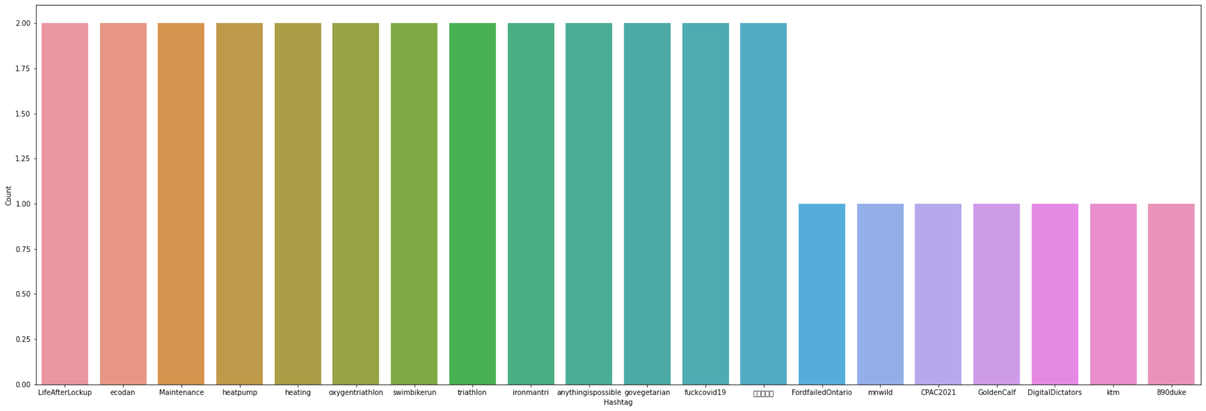


Most hashtags used in the tweets were:

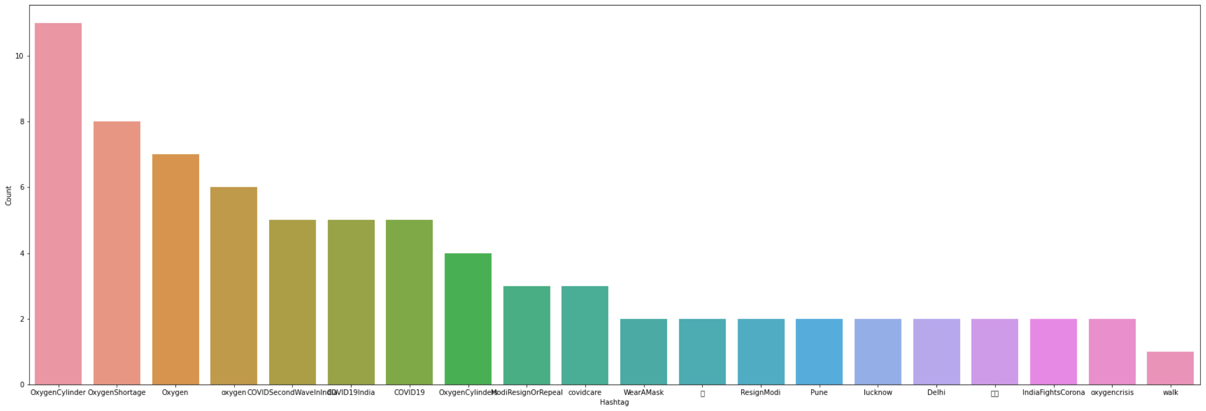
1. Help seeking tweets :



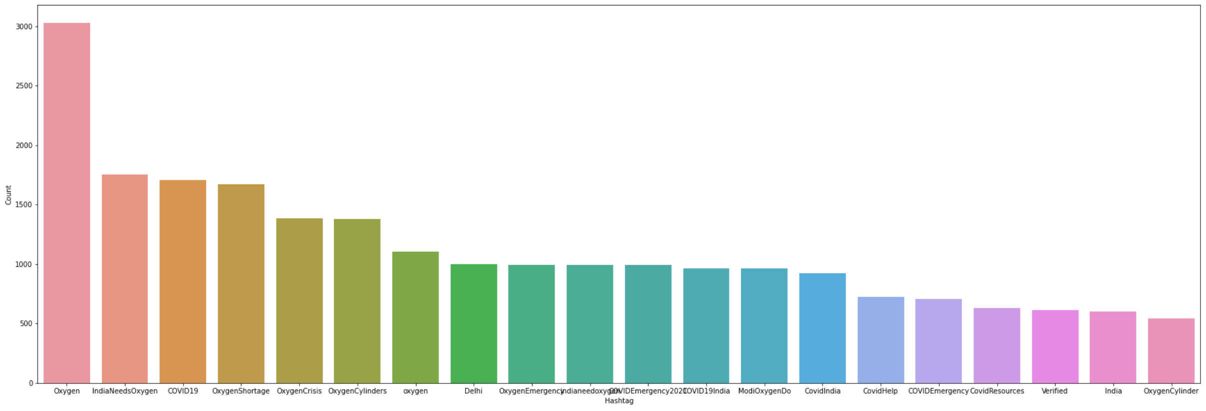
2. Tweets which are not help seeking :



3. Tweets which are neutral :



4. Most used hashtags in general in all tweets :



Insights on data :

1. Most used words in general of all tweets are Oxygen, OxygenShortage, OxygenCrisis, OxygenEmergency etc. all related to oxygen.
2. Most used words in help seeking tweets are also related mostly Oxygen, Oxygen Cylinder , covid and remedisivir, Sonusood etc.

3. Most words used in not help seeking tweets contain words which are not related to oxygen or covid and contain random words like LifeAfterlockdown, maintenance, heating etc.Neutral words also contain most words related to oxygen.
4. Since there is no major difference between neutral and positive tweets hashtags , we cannot use them to label unlabelled data.

Train set	Contains the 75% of labelled data
Validation set	Contain the 25% of labelled data
Test Set	Contain all unlabeled data

Methodology

Since data contains mostly unlabeled data and only very less labelled data (less than 1%) it is difficult to train the classifier.So what we do , we will use semi- supervised method in which we will first train the model using only labelled data and then using it we will predict the unlabelled data and and then we will again train the model but with both labelled and unlabbeled data . Using this mechanism and repeating it many times we can improve the training of our model.But repeating training of 70k data can be easily done for simple models like Naive bayes and Xgboost but for big models like bert it is not possible . So for that we will train only once with the labelled data to see how much accuracy score they give .All other models are trained once on both labelled and unlabelled data.

Tokenization

- ### WordVec

  We have tokenized every tweet in a column of some 200 features that has been taken from the corpus of whole tweets. It will be used for XGBClassifier, SVM and Random Forest model. In it we have trained the wordvec model from our data of 70k tweets only.

- ### BERT Tokenization

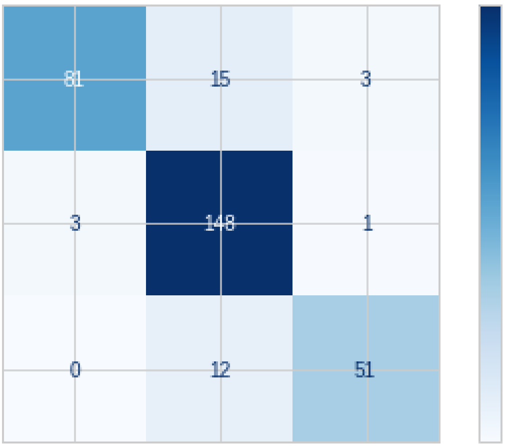
  In this we have tokenized the tweets from pretrained tokenizers available that have been trained on very large data. It will be used for the BERT model only.Here we are using the pretrained model since training the BERT model on our data will take a lot of time and resources.

Semi-Supervised Models

- ## SVM Model

First of all , we use an SVM classifier to train on labelled data and then using it predict the unlabelled data and then retrain the model using that both labelled data and predicted unlabelled data. SVM is an ideal choice since it is good to classify the data.

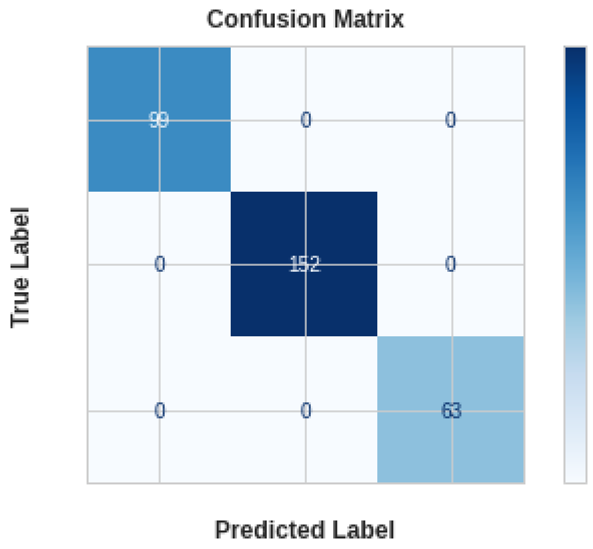
Now after training by both labelled and unlabelled data , it gives the accuracy score of 0.82 and 0.88 on validation set and labelled data respectively.



- ## Random Forest

Random Forest is also one of the good models to solve the classification problems. Same as the previous model we will first train labelled data and then predict unlabelled data and then retrain using both labelled and predicted unlabelled data with trained labels.

After training on the data twice[including unlabelled data] , the random forest gives the accuracy score of 0.96 on labelled data and 0.68 on validation set which shows its better model than SVM and XGBClassifier. It also has the fear of getting overfit.



- ## XGBClassifier

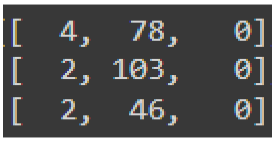
Now we will try the same with XGBClassifier and see what it scores on the data.

- ## BERT

Now we use the BERT classifier which is the one of the best transformer based ML models which can classify our tweets in the most accurate way. It is the model by google which is trained on large amount of data and we can download it and tokenize the our tweets with it and then trained the pretrained model on our data so that it can take pattern in our tweets and then through that we can predict the the unlabelled data and then again train the data with both labelled and unlabeled data.

But since it is a heavy model we are able to train the data once since the labelled data was less and then unlabeled data due to huge data was not able to predict due to huge data and limitation of the machine.

Validation accuracy = 0.5443
Accuracy on labelled set after training = 0.45531
Confusion Matrix on labelled data(this shows data is not overfit)



```
[  4,  78,   0]
[  2, 103,   0]
[  2,  46,   0]
```

Note - These scores are after training only on labelled data and not by mixing both labelled data or unlabelled data. Maybe after that its score will rise. But I cannot train that model due to high amount of data and long training time of BERT model

Result

Model	Validation Score	Labelled data score
SVM	0.82278	0.880851
Random Forest Classifier	0.745762711864406	1.0
XGB classifier	0.68182	0.96969697
BERT	0.5443	0.45531

Conclusion

So as we can see the best model among all is Random Forest Classifier as it gives a Validation score of 0.7457 . It can have the problem of overfitting on labelled data which can be reduced with the help of introducing more labelled data. Among other models all give good validation scores except BERT. Maybe this can be due to BERT trained only once on labelled data.We can train more than once to see the change. **Also we have actually taken a pretrained model of BERT which is trained in different databases(USA region) due to which it is not familiarised with Indian words like political acronyms and other due to which it performs bad on this data**. We have used Semi-supervised mode in which we train once on labelled data and then using that we predict unlabelled data and use both labelled and unlabelled data.We have trained only once after joining both data here due to machine restriction but we can train many times to increase the accuracy of the model.

Scope of improvement -

1. We can use hashtags to label some unlabelled data.
2. We can make some models to take emojis also to take out information from tweets.
3. We can trained tokenization model of BERT from the 70k tweets data so that it will familiarise with most of the indian vocabullary.

What i learned -

1. I learned to work on many different libraries like huggin face, regex, spacy, wordcount etc.
2. I learned to make systematic and full ml pipelines.
3. Learned more about semi- supervised learning.