Drug reviews

SENTIMENT ANALYSIS CLASSIFICATION

Cristina Porta Serrano

Index

- Motivation and goals
- Data and analysis
- Methodology
    -Approach
    -Model and results

Motivation and goals

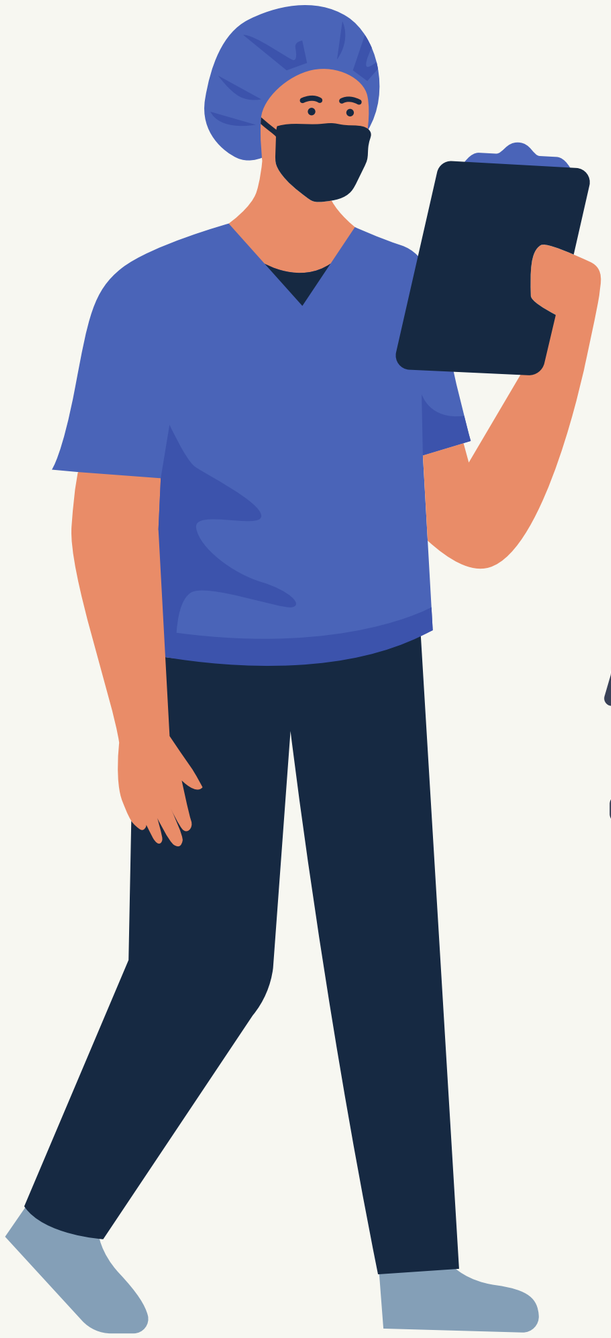
Motivation and goals

PROBLEM

My cat was having severe crisis of seizures due to an epilepsy caused by a brain injury. Nowadays, there are no epilepsy veterinary drugs, so we have to use human drugs and we want to know which is the best one.

SOLUTION

Automatize an artifical intelligence algorithm to determine the sentiment of the drug reviews so it's easier to decide which one to use.



Analysis of the data

Data

The dataset was originally published on the UCI Machine Learning repository.

The dataset provides patient reviews on specific medications along with related conditions and a 10-star patient rating that reflects overall patient satisfaction. Data was obtained by crawling online pharmaceutical review sites.

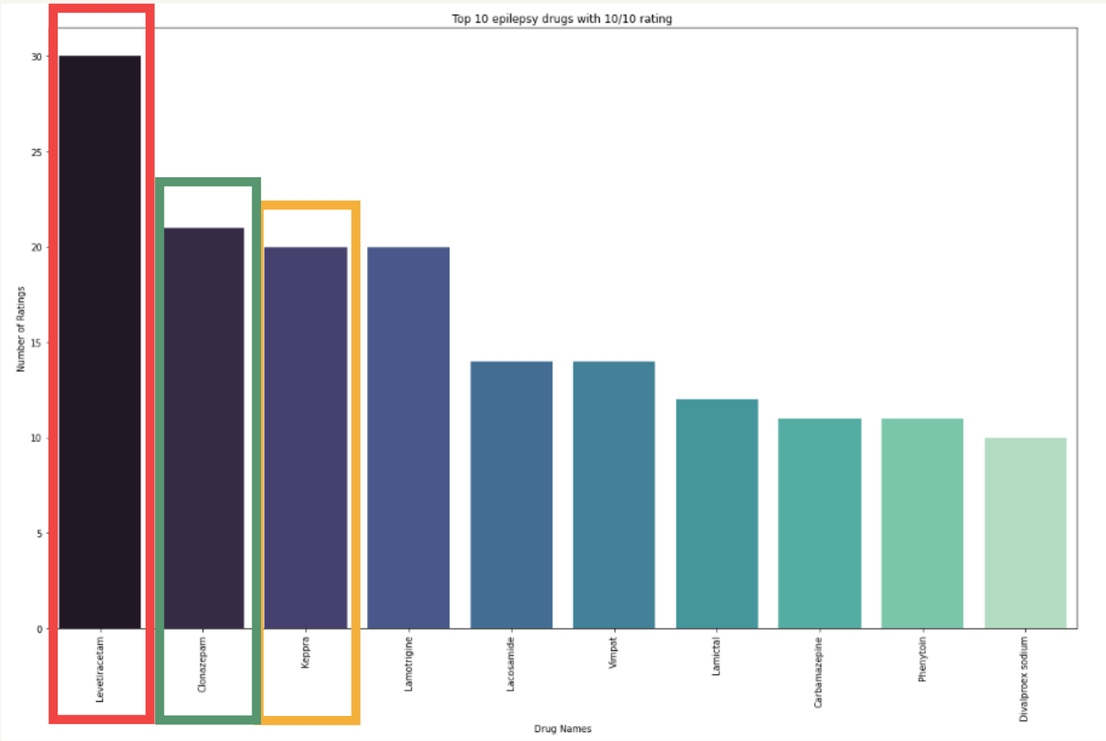
Attribute information:

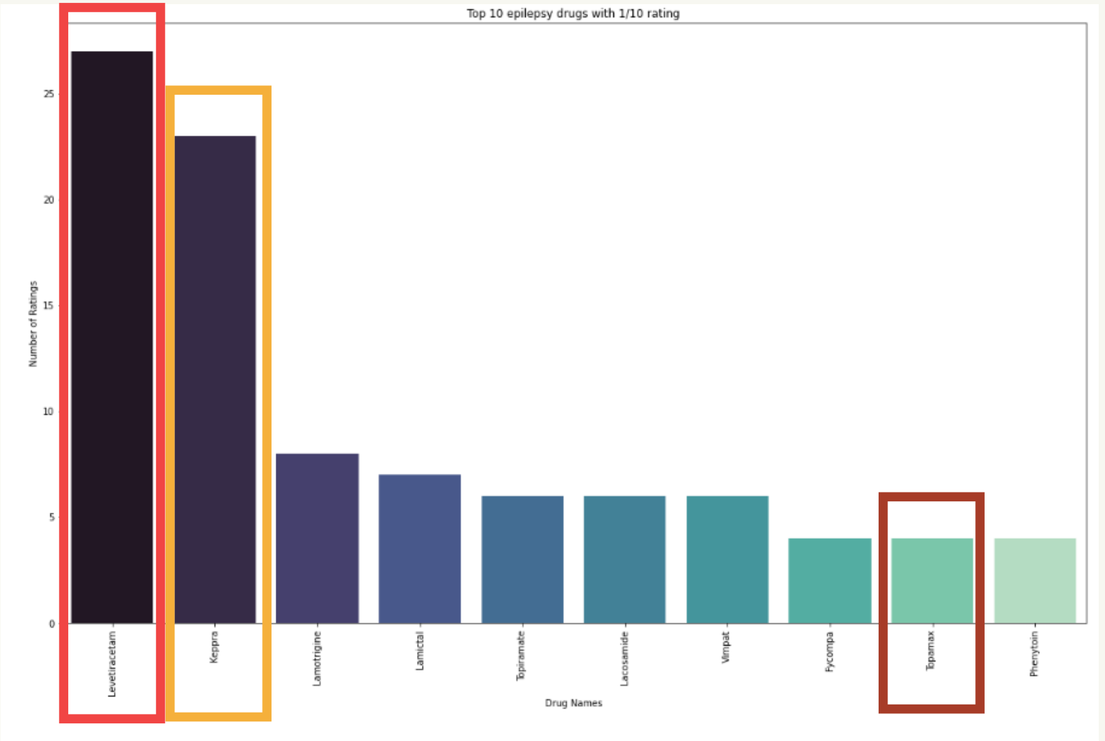
1. drugName (categorical): name of the drug.
2. condition (categorical): name of the condition.
3. review (text): patient review.
4. rating (numeric): 10-star patient rating.
5. date (date): revision entry date.
6. usefulCount (numeric): number of users who found the review useful.

Data analysis

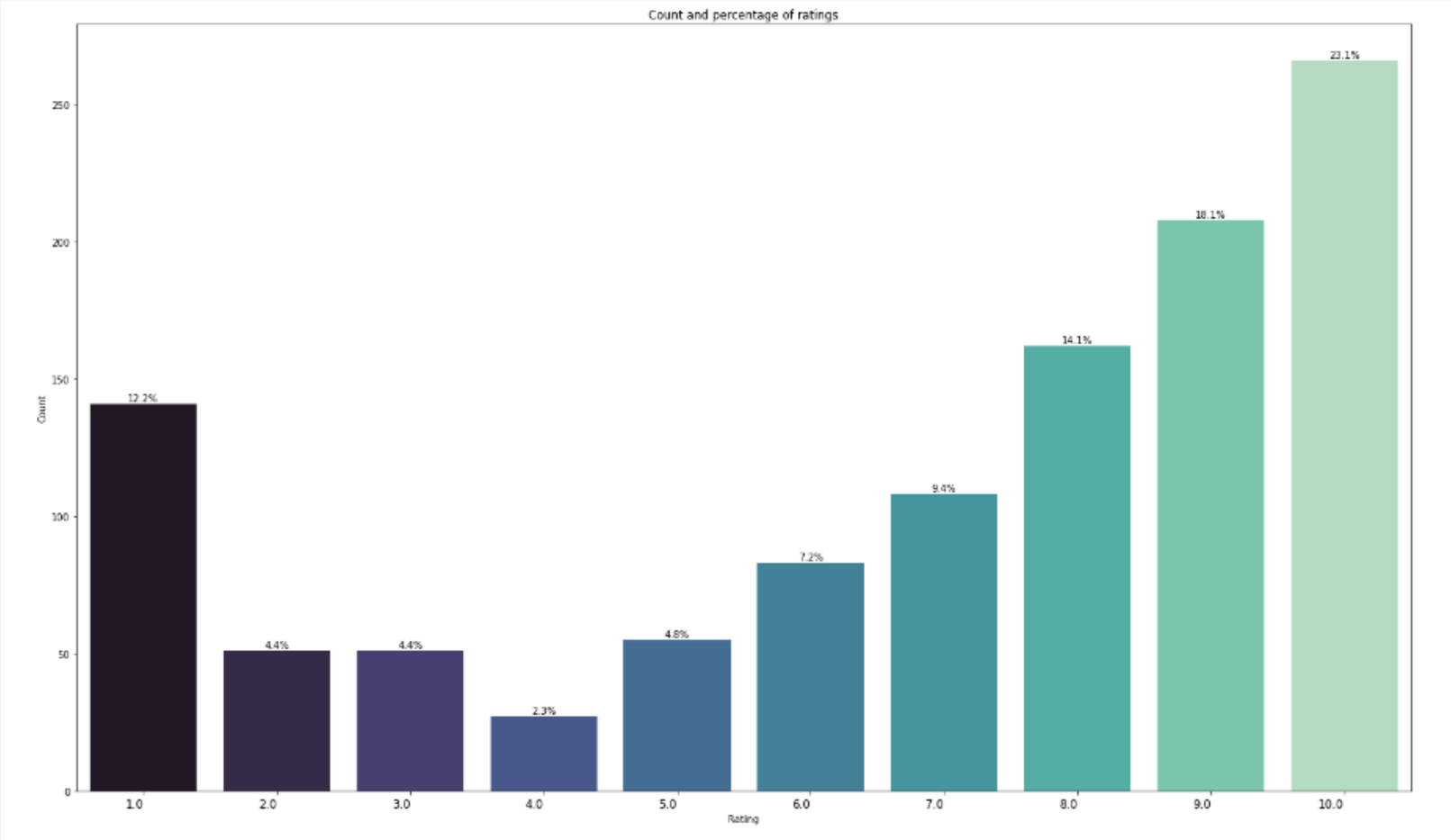
Levetiracetam is the most used drug for epilepsy/seizures so it is the one with highest and lowest ratings at the same time. Keppra is also a very used drug so we have the same problem. The drug with most 10 rating and without low rating is Clonazepam.

As the drug with most low ratings and no 10 ratings we have Topamax.





Most of the drugs used for epilepsy/seizures have good ratings. More than 55% of the ratings are 8/9 and 10. We can see that there is a 12% of 1 ratings too, so the distribution is a little bit skewed to both sides, we don't have many ratings from 2 to 6.



We replaced the rating column with the sentiment one in order to have the class label to perform our prediction of sentiment analysis (class 1 is positive and class 0 is negative). As we have seen before, most of the sentiments are positive towards the drugs used for epilepsy and seizures.

This means we have IMBALANCED DATA.

condition: seizures

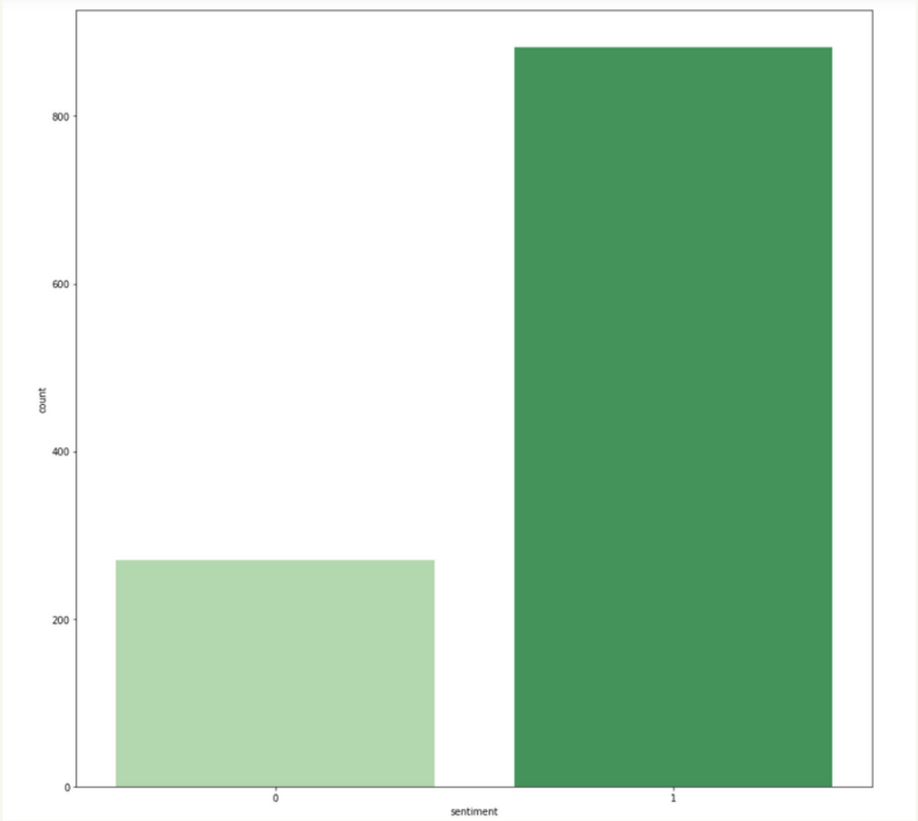
1	431
0	125

condition: epilepsy

1	283
0	100

condition: seizure prevention

1	168
0	45



Here we can see how the actual reviews are and how we cleaned them (without stopwords, stemming...)

```
drugs_sent_int[drugs_sent_int.condition == 'seizures']['review'].iloc[0]
```

```
'"having tried 7 other medications none of which controlled my seizures i was happily surprised that a dose of 4mg of fycompa d
oes. more than 12 months and no seizures. luckily i have not experienced suicidal thoughts. the only side effects i experience
are a little aggression and tiredness but only very mild."'
```

```
drugs_sent_int[drugs_sent_int.condition == 'seizures']['review'].iloc[5]
```

```
'"had a brain tumor removed in 1988 and was on dilantin and phenobarbital for over 20 years. got rushed to the hospital for wha
t was thought to be a stroke and ended up being a seizure. my new nuero dr. put me on keppra with phenobarbital, this worked gr
eat for about 5 years. just got out of the hospital again because of a seizure and am now on 3,000 mg of keppra and 400 mg of z
onisamide daily. at first, i had a hard time sleeping,only about an hour a day, my fault, as i read the prescription bottle wro
ng. i was suppose to take the zonisamide once in the morning and here i was taking 200 mg in the morning and 200 mg at night. s
o far everything is fine as i am only on week 1 of my new dosage. my dr. also wants me to lose weight, so this is it."'
```

```
drugs_sent_int[drugs_sent_int.condition == 'seizures']['review_clean'].iloc[0]
```

```
'tri 7 medic none control seizur happili surpris dose 4mg fycompa 12 month seizur luckili experienc suicid thought side effect
experi littl aggress tired mild'
```

```
drugs_sent_int[drugs_sent_int.condition == 'seizures']['review_clean'].iloc[5]
```

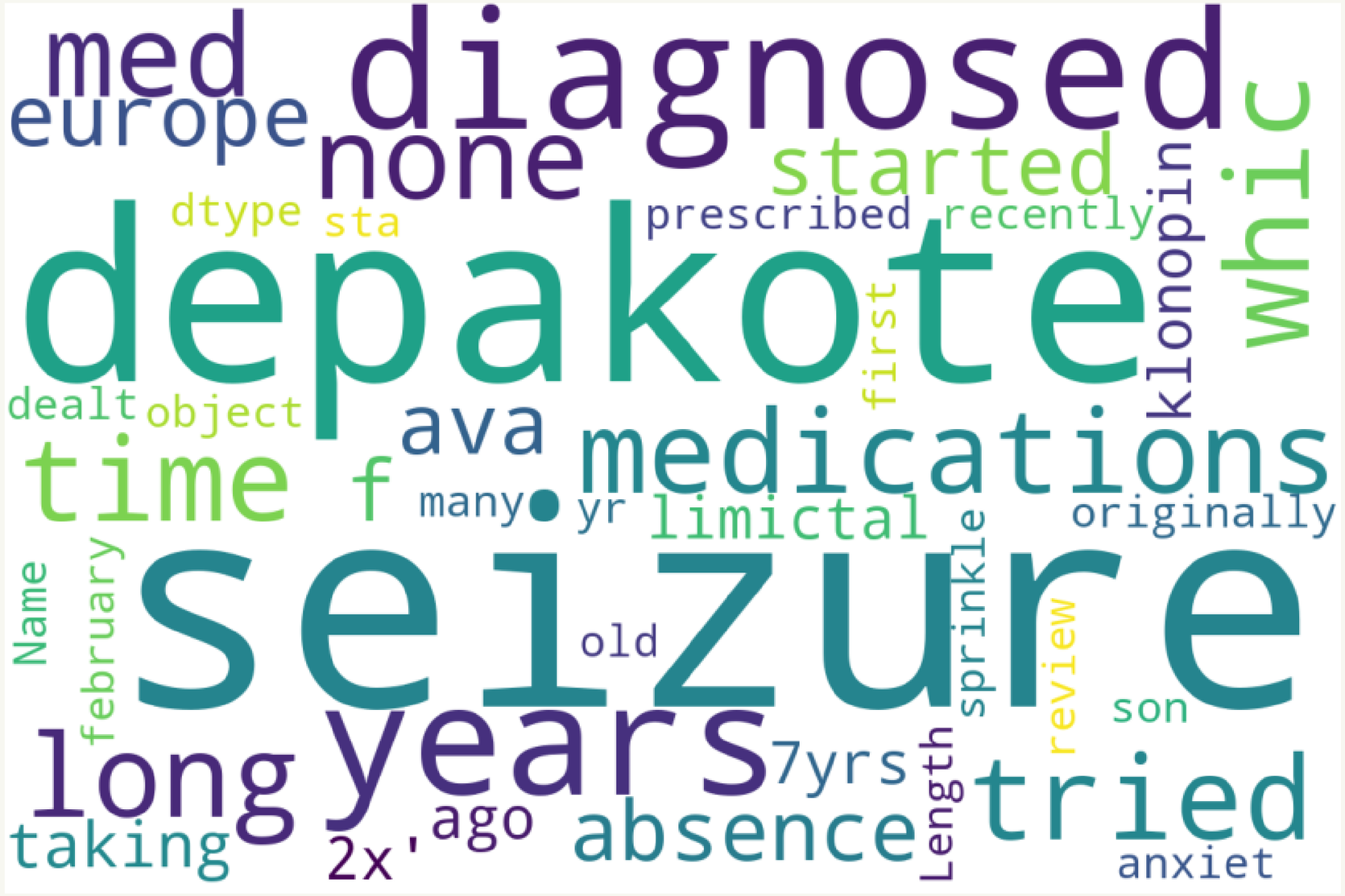
```
'brain tumor remov 1988 dilantin phenobarbit 20 year got rush hospit thought stroke end seizur new nuero dr put keppra phenobar
bit work great 5 year got hospit seizur 3 000 mg keppra 400 mg zonisamid daili first hard time sleep hour day fault read prescr
ipt bottl wrong suppos take zonisamid morn take 200 mg morn 200 mg night far everyth fine week 1 new dosag dr also want lose we
ight'
```

Methodology

How do we approach it?

OUR VISION OF THE PROJECT

We want to create the best prediction model. First of all, we clean the data: search for NaNs to deal with them if any (in our case there weren't NaNs), remove unnecessary columns and create new interesting ones with information that can help us develop the best model. We use some predefined libraries to create them. Then, we visualize the data in graphs to have a better sight of it.

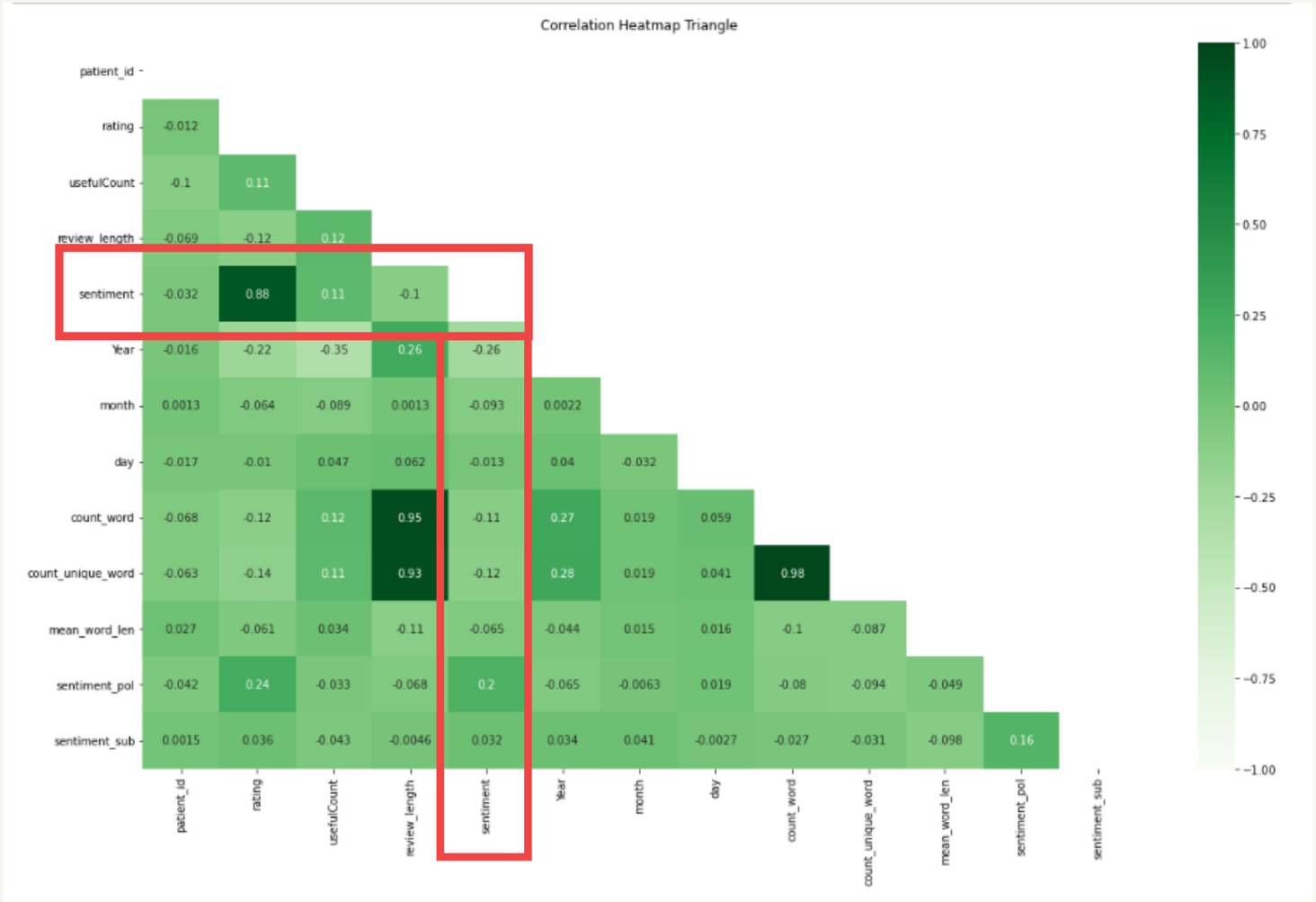

med
diagnosed
europe
none
started
depakote
dtype
sta
prescribed
recently
klonopin
whic
dealt
object
ava
first
time
f
medications
many
yr
limictal
originally
seizure
Name
february
old
sprinkle
review
son
years
long
7yrs
tried
taking
2x
ago
absence
Length
anxiet

Don't forget that what will give us the predictions are the reviews, so we are dealing with text. We need to clean it:

- changing to lower case
- replacing the repeating pattern of &#039
- removing all the special characters
- removing all the non ASCII characters
- removing the leading and trailing whitespaces
- replacing multiple spaces with single space
- replacing two or more dots with one
- removing the stopwords
- stemming the words

As we said, we added the sentiment column (we can see, obviously, that is very related to the rating one so we drop it). We also added the word count in each review, unique word count, average length of the words. We used the Vader Sentiment Analyzer library to add columns related to the sentiment (positive, negative and neutral) and the TextBlob one to add the sentiment polarity and subjectivity.

All these columns will help the models to perform in a better way. In the correlation matrix we see that the highest correlations are the obvious ones so we don't get much information.



Models and results

In brief, we tried these models to select the best performing one:

Traditional ML:

- Random Forest
- LGBM
- Cat Boost

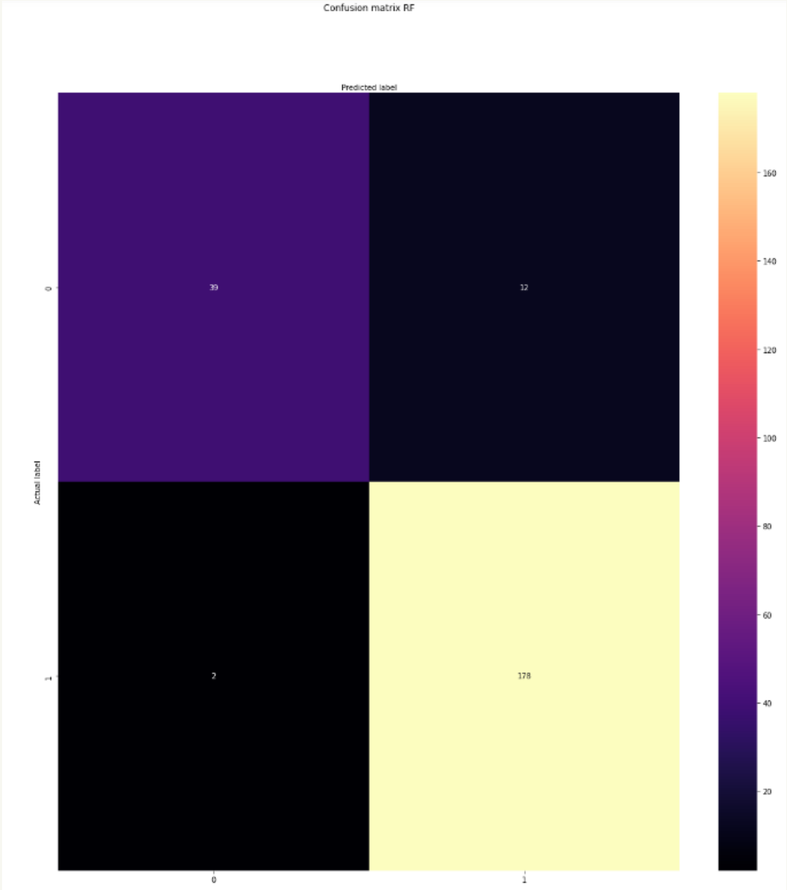
DL:

- LSTM with custom embedding
- Transfer learning with GloVe

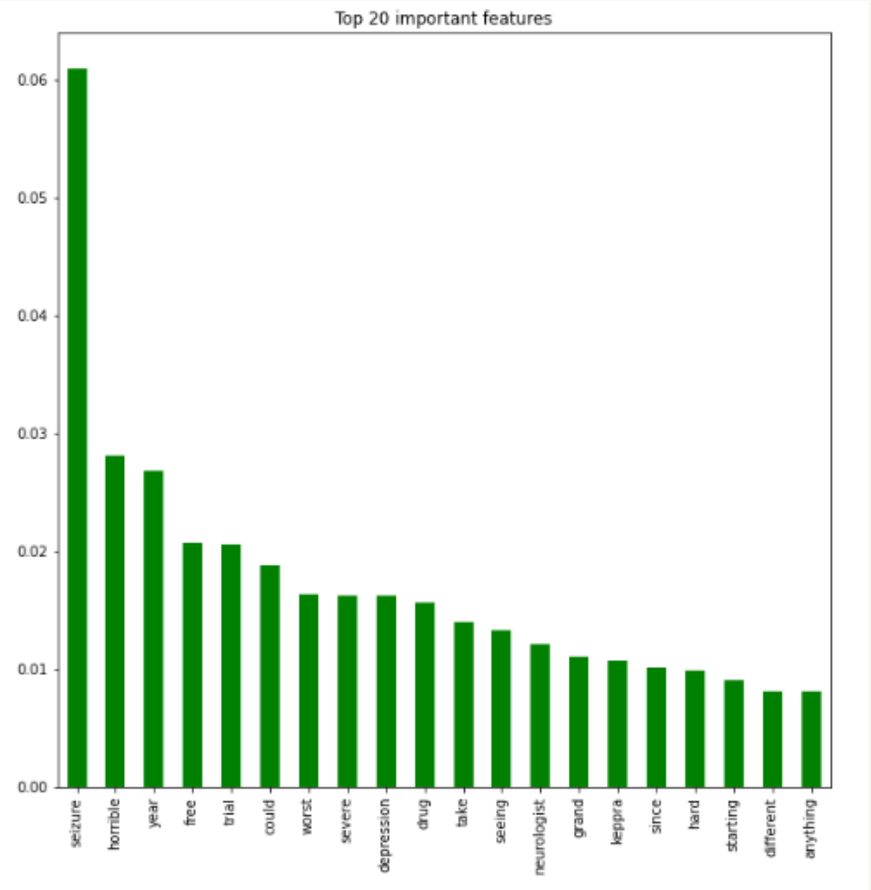
As we know, models don't work with text or images, they need numerical data. To achieve this, we need to vectorize the reviews with tdidf methodology before training the models.

Model	Dataset	Accuracy
RF	whole	80.4
RF	epilepsy	93.9
LGBM	whole	85.4
LGBM	epilepsy	92.2
CB	whole	81.3
CB	epilepsy	91.7
LSTM custom 1	whole	80.8
LSTM custom 1	epilepsy	77.9
LSTM pretrained	whole	79.2
LSTM pretrained	epilepsy	87

Confusion matrix RF



Feature importance RF



Thank you for your attention!