Interpretable & Explainable AI - FAIP Week 7 - Assignment

Context

Being able to interpret and explain the current generation of AI systems is a fundamental challenge. In this assignment, you will train a Random Forest classifier on task of Sentiment Analysis and produce explanations for its predictions.

To explain the classifier, you will be using LIME. You can refer to the video lectures and the original paper for details on how it works.

Data

The dataset you will be using is made of Reddit comments together with the respective sentiment labels. The CSV file containing the data can be downloaded from this assignment's page on Brightspace.

For the purpose of this assignment, consider those labels to be correct.

Tasks for this Assignment

- **Task #1**: Describe when, and if, the explanations produced using LIME seem accurate (or not), and why is that the case
- **Task #2**: Design a human evaluation task to assess the quality of the explanations you obtained

Once you are ready to submit the assignment, export the notebook as PDF and upload it on Brightspace.

Task #1 - Generate and Describe Explanations

Installation and Package Import

In [ ]:
```
# == Do not change this ==
# !pip install lime
```

In [1]:
```
import pandas as pd

# Text processing inputs
from sklearn.pipeline import make_pipeline
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_extraction.text import CountVectorizer

# LIME Explainer
from lime import lime_text
from lime.lime_text import LimeTextExplainer
```

Model Training

Load data

In [2]:
```
df = pd.read_csv("reddit_clean.csv", sep=";")
df.head()
```

Out[2]:

	category	comment
0	1	family mormon have never tried explain them th...
1	1	buddhism has very much lot compatible with chr...
2	-1	seriously don say thing first all they won get...
3	0	what you have learned yours and only yours wha...
4	1	for your own benefit you may want read living ...

Prepare data

In [3]:
```
cv = CountVectorizer(binary=True, stop_words='english')
cv.fit(df.comment)
```

```
X = cv.transform(df.comment)
y_train = df.category
```

Create model

In [4]:
```
rf = RandomForestClassifier(n_estimators=200)
rf.fit(X, y_train)
```

Out[4]:
```
RandomForestClassifier(n_estimators=200)
```

Adding LIME to the pipeline

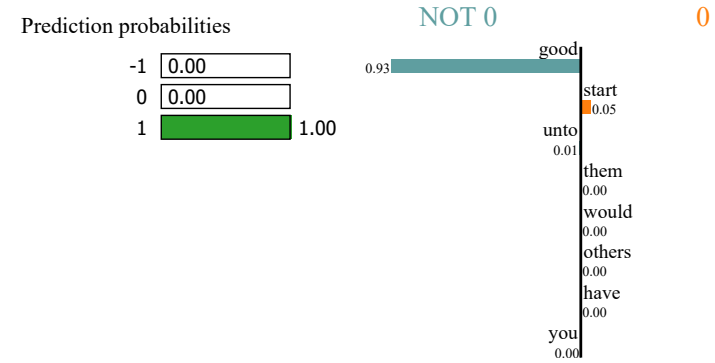
In [5]:
```
c = make_pipeline(cv, rf)
explainer = LimeTextExplainer(class_names=[-1,0,1])
```

Generate explanations with LIME

- Change the value of the variable `idx` to pick different Reddit comments

In [8]:
```
idx = 10
exp = explainer.explain_instance(df.comment[idx], c.predict_proba, num_features=8)
print('True class: %s' % df.category[idx])
exp.show_in_notebook(text=True)
```

```
True class: 1
```



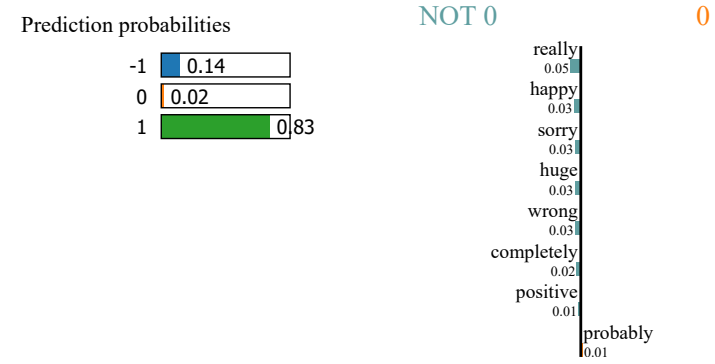
Text with highlighted words

unto others you would have them unto you would good start

This is a good explanation, because it picks a word in the positive classification to be the predicted sentiment.

In [18]:
```
idx = 6
exp = explainer.explain_instance(df.comment[idx], c.predict_proba, num_features=8)
print('True class: %s' % df.category[idx])
exp.show_in_notebook(text=True)
```

```
True class: 1
```



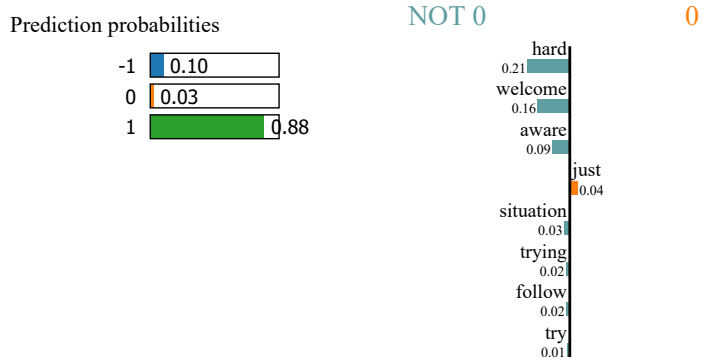
Text with highlighted words

was teens when discovered zen meditation was then undiagnosed bpd being homeschooled and just gotten 56k modem with web connection where came across link zen meditation tried for couple weeks and the change was palpable felt the most profound sense peace ever felt grades immediately started going had more energy started martial arts just huge positive change all around parents asked something was wrong fundie parents when anything changes this was where naiveté kicked foolishly told them been trying meditation and really calmed down thought they happy that

This is not a good example, because you don't really see why the model chose 1 to be the predicted probability (is "really" positive or ngeative?)

In [19]:
```
idx = 8
exp = explainer.explain_instance(df.comment[idx], c.predict_proba, num_features=8)
print('True class: %s' % df.category[idx])
exp.show_in_notebook(text=True)
```

```
True class: 1
```



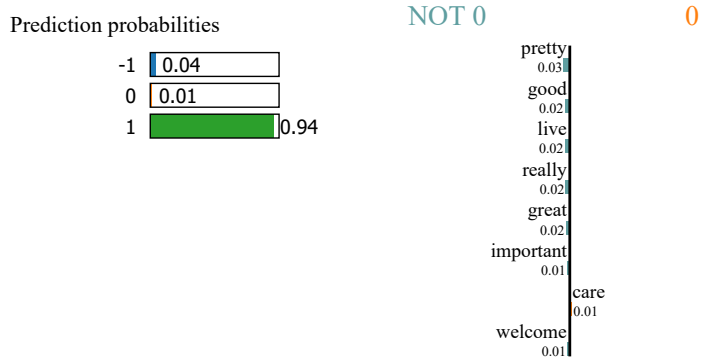
Text with highlighted words

dont worry about trying explain yourself just meditate regularly and try hard you can more aware everything else will follow coming from someone who has been throught his situation welcome pms

This is a good example, because it picked the right words to be predicted as positive.

In [20]:
```
idx = 9
exp = explainer.explain_instance(df.comment[idx], c.predict_proba, num_features=8)
print('True class: %s' % df.category[idx])
exp.show_in_notebook(text=True)
```

```
True class: 1
```



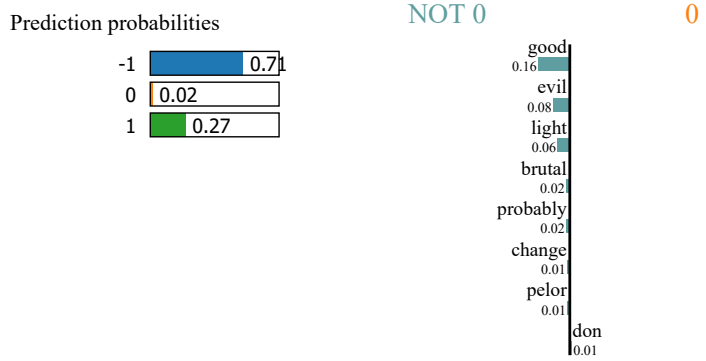
Text with highlighted words

recently told family that buddhist live the bible belt this whole ordeal involved leaving the baptist church and everything been pretty rough but those who really care about have been open and accepting they seen the good has created life and relationships with others fact there are handful christians who have lovely conversations with and that truly respect someone else suggested living buddha living christ great one read about the important dialogue between buddhists and christians also welcome you message

Most of the positive words are being detected to be positive. However, the model also missed a few positive words ("lovely", "truly") who could also be helpful for classifying this post as positive.

In [23]:
```
idx = 19
exp = explainer.explain_instance(df.comment[idx], c.predict_proba, num_features=8)
print('True class: %s' % df.category[idx])
exp.show_in_notebook(text=True)
```

```
True class: -1
```



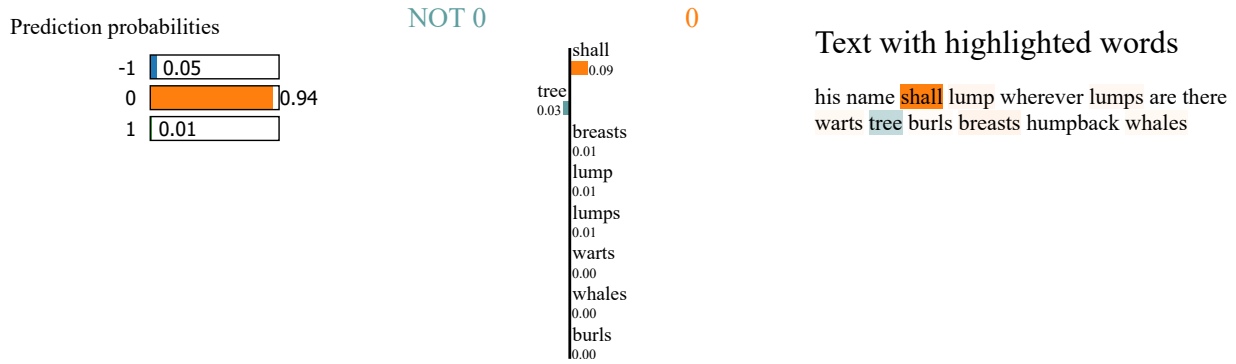
Text with highlighted words

though don have any good suggestions for additional deities the moment you should probably have some overlap the cavern dwelling gnolls might consider pelor the god evil blinding them with his wicked light loki has sufficiently fooled them over the years into thinking their benefactor always leading them the bountiful crops they steal from the peasantry they have statues him fat woman giving out food the starving they view the god death the god justice giving righteous end those gnolls that have served brutal conflicts point view can change everything

This is a bad example, a lot of words are missed for the negative prediction. And "good" in this context can be linked to "don" which is not-good instead of good. So this is also labelled wrongly.

In [24]:
```
idx = 23
exp = explainer.explain_instance(df.comment[idx], c.predict_proba, num_features=8)
print('True class: %s' % df.category[idx])
exp.show_in_notebook(text=True)
```

```
True class: 0
```



This is a good example, because there are no identifiable words to be predicted as positive or negative.

Copy the cell above to show multiple (good and bad) examples ...

Task #2: Design Explanation Human Evaluation

Now that you have seen the kind of explanations LIME produces, assume you have a collection of them that need to be evaluated.

Based on the explanation evaluation criteria discussed in the video lectures, design an evaluation task involving humans.

Things you need to consider are:

- Properties of explanations
- How to present the explanations to people
- Do they need any additional information to understand them?
- Level at which you want the explanations to be evaluated

Motivate your choices.

Hint - when designing this, you can refer to the concepts discussed in Data Work 1 & 2 (Week 2).

Task Design

Aim of evaluation: how well these explanations align with human understanding of sentiment and how useful and reliable/trustworthy the explanations are to end users. With this, humans can evaluate and improve models.

Properties of explanations Things that are good to be considered (in a survey or validation/verification) are as follows:

- Fidelity: does the model explanation accurately represent the model's reasoning? (highlighted parts contribute to model's prediction)
- Fidelity: are all relevant aspects covered, or are there missing parts that need to be covered?
  - These questions are relevant to how useful and reliable the explanations are.
- Intelligibility: is the explanation easy to understand for humans? i.e. Do the highlighted parts clearly show why the sentiment is classified in a certain way?
- Intelligibility: is the explanation aligning with human's classification of sentiment?
  - These questions are relevant for the comprehensiblity of the interpretation to be presented to humans

Presenting the explanations to people In the task itself, each participant will be presented with

- the original reddit post
- the true prediction
- the LIME explanation as presented above (models prediction, LIME explanation)

Additional information

- description of task procedure
- a short explanation will be provided with some practice trials
- maybe provide them with some examples of good and bad explanations

Level evaluations This task is including real reddit posts and involving real humans, so this can be described as application-grounded evaluation. If the task is a more simplified version of a real sentiment analysis, this would be a human-grounded evaluation. Here, it is most important that there is human involvement to assess whether this the explanation is accurate and easy to understand for humans.