There is NO official lrp package than can be used.

grad-cam can in fact be used with transformers with the "official" package.

grad-cam can be used with "vision" transformers with the "official" package.

grad-cam can't be used with classical ml

Lime and SHAP don't care about the type of model.

grad-cam works only on pytorch

https://github.com/albermax/innvestigate is an lrp implementation using keras

https://github.com/chr5tphr/zennit is an lrp implementation using pytorch

https://github.com/jacobgil/pytorch-grad-cam is a grad-cam pytorch implementation

Both lrp and grad-cam have relatively complete implementation for CNNs

Both can be used in NLP but with CNN models only (no attention-based or transformer-based models)

Both have incomplete implementation in case of attention-based or transformer-based models

Both LRP and grad-cam CAN be used with attention-based or transformer-based models but we will need to implement them from scratch

logistic regression need much more iterations than the default value to successfully converge (now it is set to 10,000 iterations maximum)

Shoutout for <https://github.com/yidinghao/interpreting-nlp/tree/master> for saving the day and enabling us to do the LRP

Embedding needs to be done in batches on GPU

logistic regression with light stemming vectorizer:

A screenshot of a computer screen

Description automatically generated

logistic regression with root stemming vectorizer:

A screenshot of a computer screen

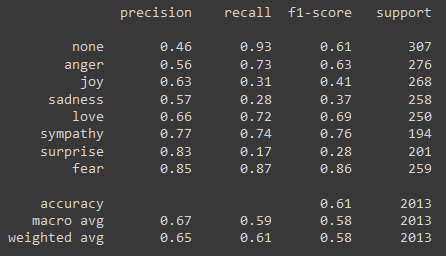
Description automatically generated

Naïve bayes with Light Stemming vectorizer:

A screenshot of a computer screen

Description automatically generated

Naïve bayes with root Stemming vectorizer:



logistic regression with light stemming embedding:

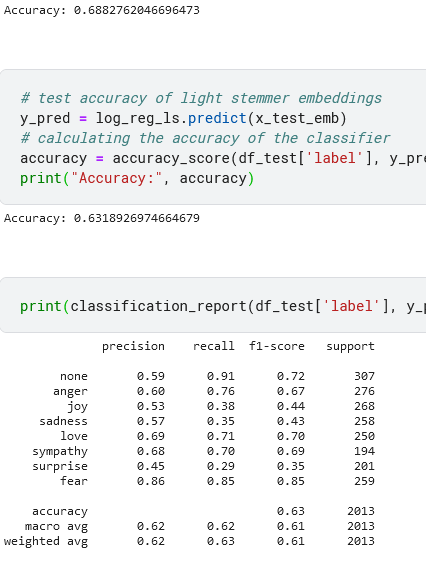
A screenshot of a computer screen

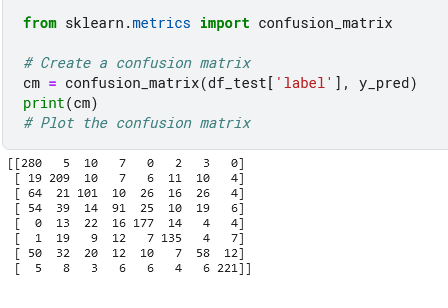
Description automatically generated

A screenshot of a computer

Description automatically generated

logistic regression with root stemming embedding:





Examples:

None predicted correctly:

A group of text on a white background

Description automatically generated

A screenshot of a computer

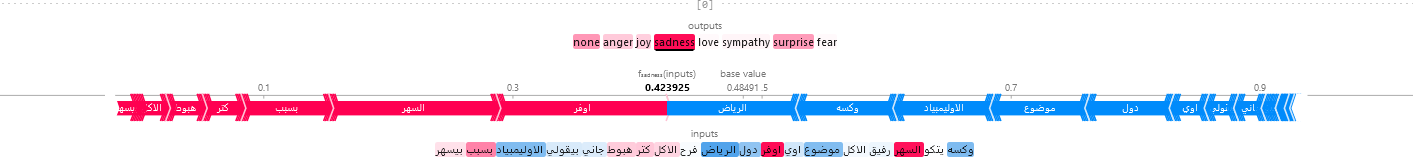
Description automatically generated

We can observe that the prediction is true, but only because of the “الاوليمبياد” word

None predicted incorrectly

A group of numbers and lines

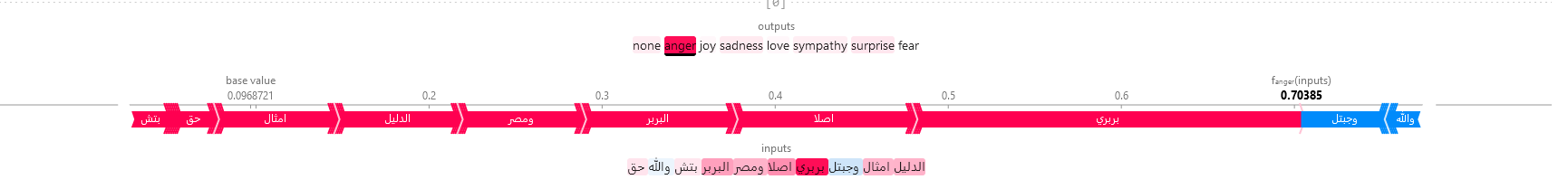
Description automatically generated with medium confidence



anger predicted correctly

A group of numbers and symbols

Description automatically generated with medium confidence



anger predicted incorrectly



A screenshot of a computer

Description automatically generated

joy predicted correctly

A group of text on a white background

Description automatically generated



joy predicted incorrectly

A white background with colorful text

Description automatically generated

A close up of a line

Description automatically generated

In this case the predicted label is different than the true label but the model, surprisingly, is not actually that wrong.

sadness predicted correctly

A white background with text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

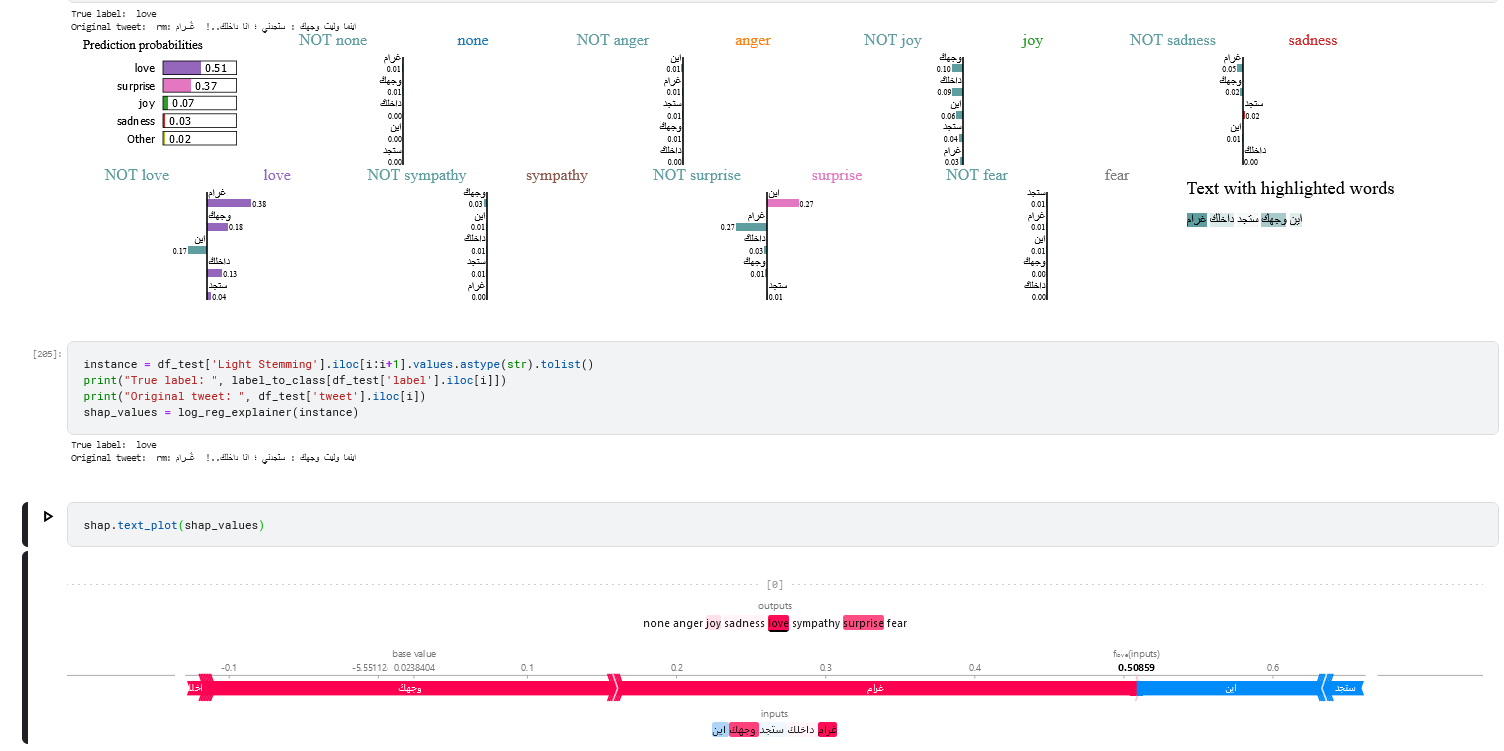
sadness predicted incorrectly

A group of numbers and symbols

Description automatically generated with medium confidence A close up of a red line

Description automatically generated

love predicted correctly



love predicted incorrectly

A group of colorful text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

sympathy predicted correctly

A group of numbers on a white background

Description automatically generated

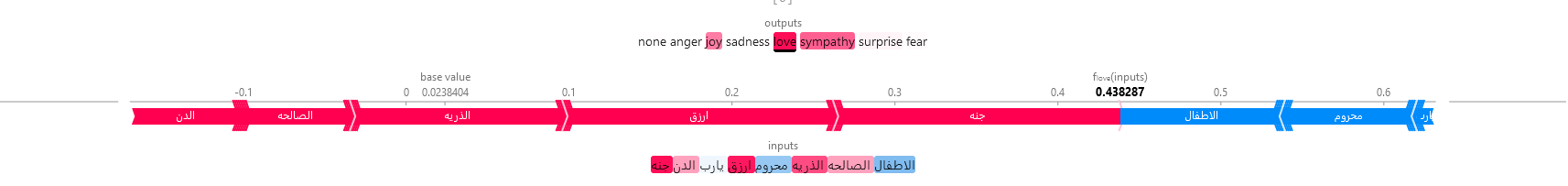
A close up of a red line

Description automatically generated

sympathy predicted incorrectly

A group of text on a white background

Description automatically generated



Surprise predicted correctly

A group of colorful text on a white background

Description automatically generated

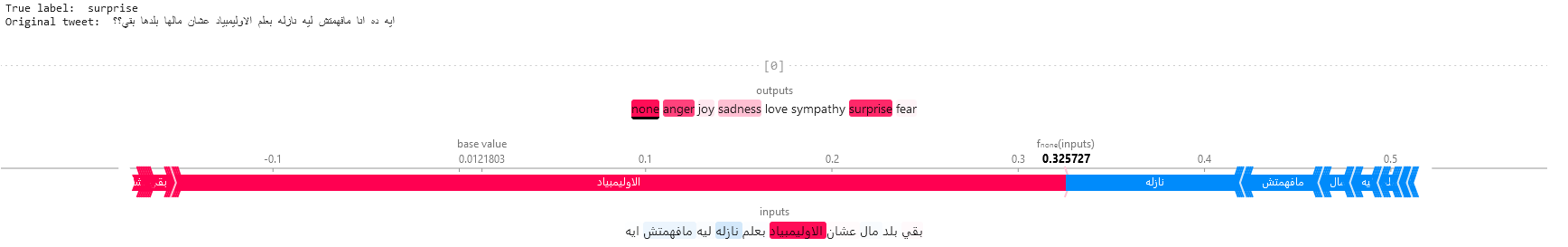
A close up of a graph

Description automatically generated

Surprise predicted incorrectly

A group of text on a white background

Description automatically generated



“الاوليمبياد” strikes once more.

Fear predicted correctly

A screenshot of a computer

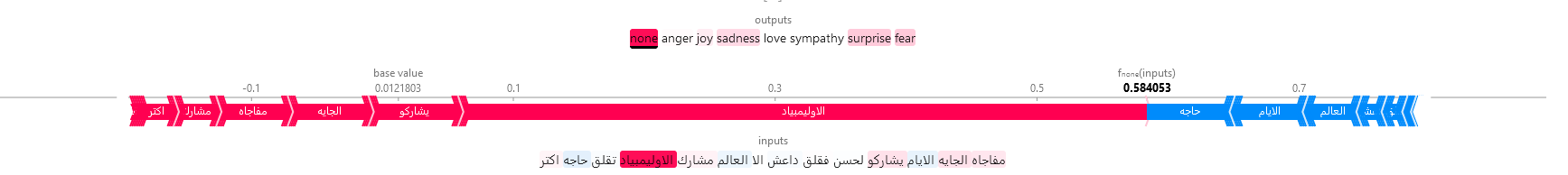
Description automatically generated

Incorrect use of emojis

Fear predicted incorrectly

A group of numbers and lines

Description automatically generated with medium confidence



Notes:

1. Light stemming is much better
2. WITH tokens is actually better than without
3. Some words(“الاوليمبياد”) mess up the prediction
4. Some emotions are close to one another and have a high chance of being confused together (ex. Sympathy and sadness, love and joy, sadness and anger, none and all)
5. Some samples are labeled incorrectly
6. Some samples’ labels ambiguous.
7. Sometimes samples are labeled correctly for weird reasons