Phishing Message Detector

As the name implies, the objective is to develop a ML model capable of detecting phishing messages. The model uses only pure text for training (no URLs or any special kind of text).

We use the model, alongside another ML model, that detects phishing URLs, made by our collegue Rafael Pinto, to detect phishing messages and URLs on discord servers. For that, we developed a discord bot.

Authors

This model was developed and trained by Pompeu Costa and João Mourão.

Training

To train the model, we used this dataset from huggingface.
Dataset credits go to its author.

Pre-Processing

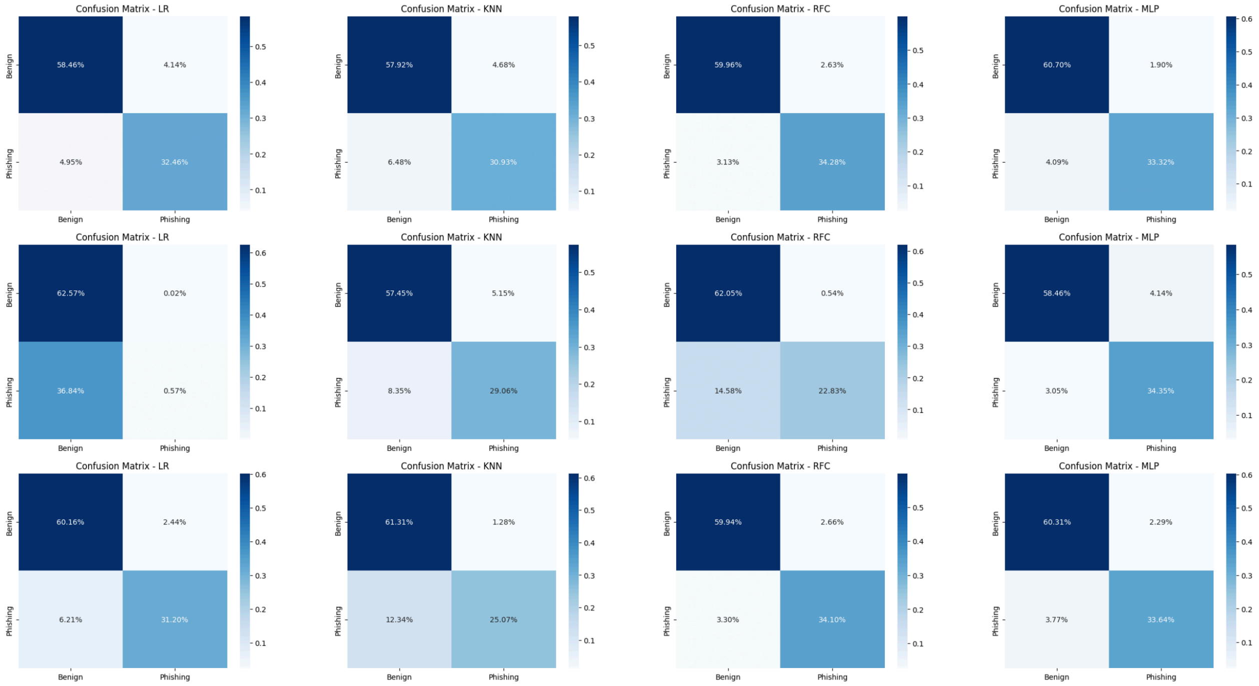
Our pre-processing consists of two steps: data pre-processing and feature extraction.

To pre-process, we just take each row of the dataset, transform it into tokens, apply lemmatization and remove stop words.
We then join the pre-processed tokens again to form the pre-processed rows.

For feature extraction, we experimented three different methods: `Bag of Words`, `Fast Text` and `TF-IDF`.
In all of the methods, we decided to use a maximum of 512 features.

The results can be seen in their respective md file: BoW, Fast Text and TF-IDF.
The parameters used for each classifier are described in the next section.

To facilitate the viewing of the results, the image below contains the results of each feature extraction method.
The columns represent the classifier and the rows represent the method (BoW -> Fast Text -> TF-IDF).



Classifiers

After pre-processing, the model needs to be trained. For that, we experimented with four classifiers: `Logistic Regression`, `K Nearest Neighbours`, `Random Forest` and `Multi-Layer Perceptron`.

To assess the best parameters for each classifier, we performed a `Grid Search` for the following parameters:

```
params = {
    'LR': {
        'C': [0.01, 0.1, 1, 10],
        'solver': ['lbfgs', 'liblinear','saga','sag'],
        'max_iter': [100, 200, 300]
    },
    'KNN': {
        'n_neighbors': [3, 5, 7, 10],
        'weights': ['uniform', 'distance'],
        'metric': ['euclidean', 'manhattan']
    },
    'RFC': {
        'n_estimators': [50, 100, 150, 200],
        'max_depth': [None, 10, 20, 30],
        'min_samples_split': [2, 5, 10],
        'min_samples_leaf': [1, 2, 4]
    },
    'MLP': {
        'hidden_layer_sizes': [(1,),(100, 100),(128,64,32), (256,128,64)],
        'activation': ['relu', 'tanh', 'identity', 'logistic'],
        'solver': ['adam', 'sgd', 'lbfgs'],
        'learning_rate_init': [0.001, 0.01],
    }
}
```

The best parameters for each classifier were the following:

- LR: {'C': 0.1, 'max_iter': 100, 'solver': 'lbfgs'}
- KNN: {'metric': 'euclidean', 'n_neighbors': 5, 'weights': 'distance'}
- RFC: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200}
- MLP: {'activation': 'relu', 'hidden_layer_sizes': (256, 128, 64), 'learning_rate_init': 0.001,'solver': 'adam'}

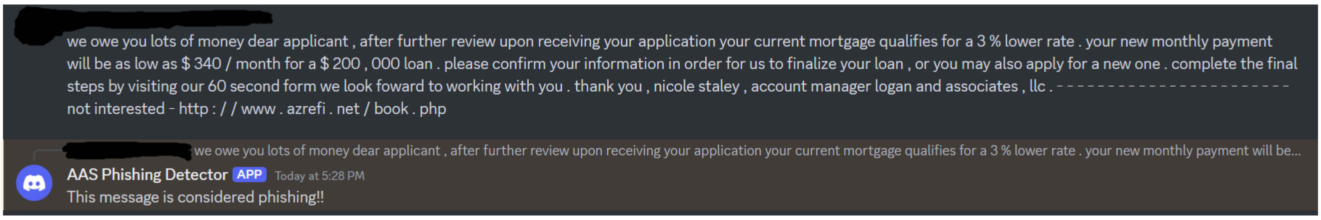
Results

Looking at the compilation image, its possible to see that the best one is the `Multi-Layer Perceptron` classifier with the `Fast Text` method.
It keeps a good accuracy while having a low rate of false positives.

Having a benign message that is considered phishing will not cause any harm, however, having a phishing message that is considered benign can cause harm to the users.
Thus, we decided to sacrifice a bit of accuracy in favour of lower false positive rates.

Running the model in the discord bot

The bot, using this model, was capable of identifying phishing messages.



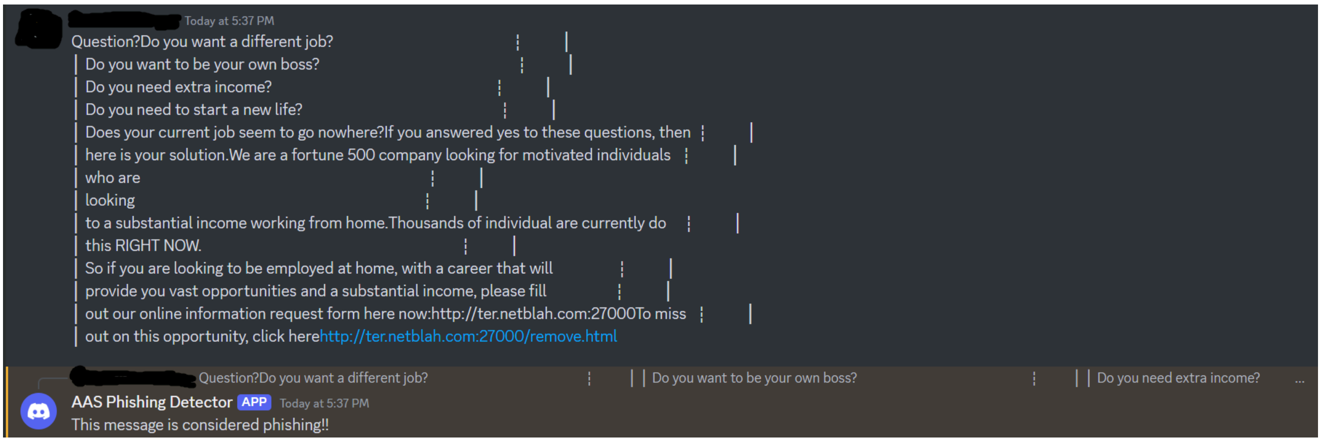
[redacted]

we owe you lots of money dear applicant , after further review upon receiving your application your current mortgage qualifies for a 3 % lower rate . your new monthly payment will be as low as $ 340 / month for a $ 200 , 000 loan . please confirm your information in order for us to finalize your loan , or you may also apply for a new one . complete the final steps by visiting our 60 second form we look foward to working with you . thank you , nicole staley , account manager logan and associates , llc . - not interested - http : / / www . azrefi . net / book . php

[redacted] we owe you lots of money dear applicant , after further review upon receiving your application your current mortgage qualifies for a 3 % lower rate . your new monthly payment will be...

AAS Phishing Detector APP Today at 5:28 PM

This message is considered phishing!!



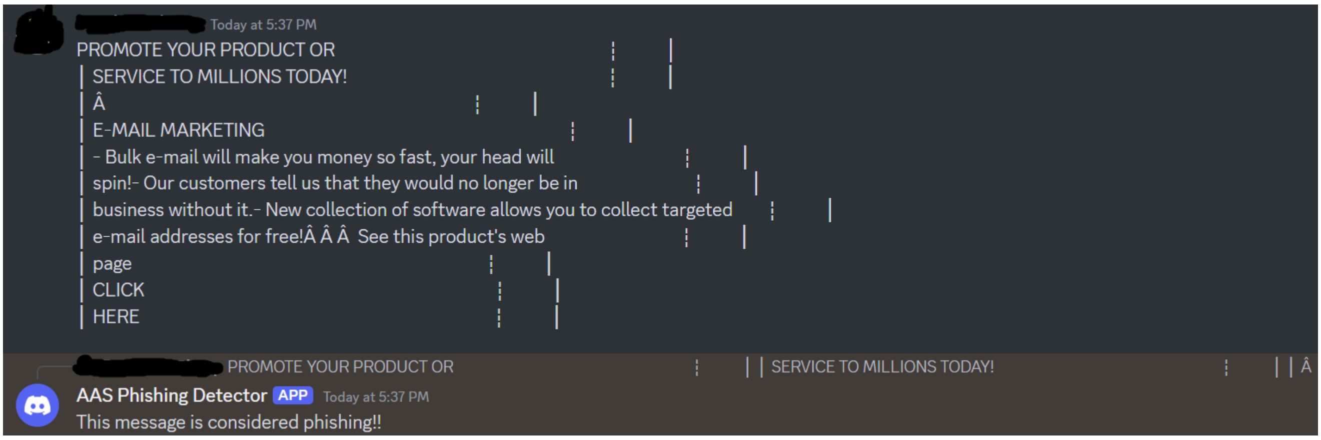
[redacted] [redacted] Today at 5:37 PM

Question?Do you want a different job? ¦ |
| Do you want to be your own boss? ¦ |
| Do you need extra income? ¦ |
| Do you need to start a new life? ¦ |
| Does your current job seem to go nowhere?If you answered yes to these questions, then ¦ |
| here is your solution.We are a fortune 500 company looking for motivated individuals ¦ |
| who are ¦ |
| looking ¦ |
| to a substantial income working from home.Thousands of individual are currently do ¦ |
| this RIGHT NOW. ¦ |
| So if you are looking to be employed at home, with a career that will ¦ |
| provide you vast opportunities and a substantial income, please fill ¦ |
| out our online information request form here now:http://ter.netblah.com:27000To miss ¦ |
| out on this opportunity, click herehttp://ter.netblah.com:27000/remove.html

[redacted] Question?Do you want a different job? ¦ | | Do you want to be your own boss? ¦ | | Do you need extra income? ...

AAS Phishing Detector APP Today at 5:37 PM

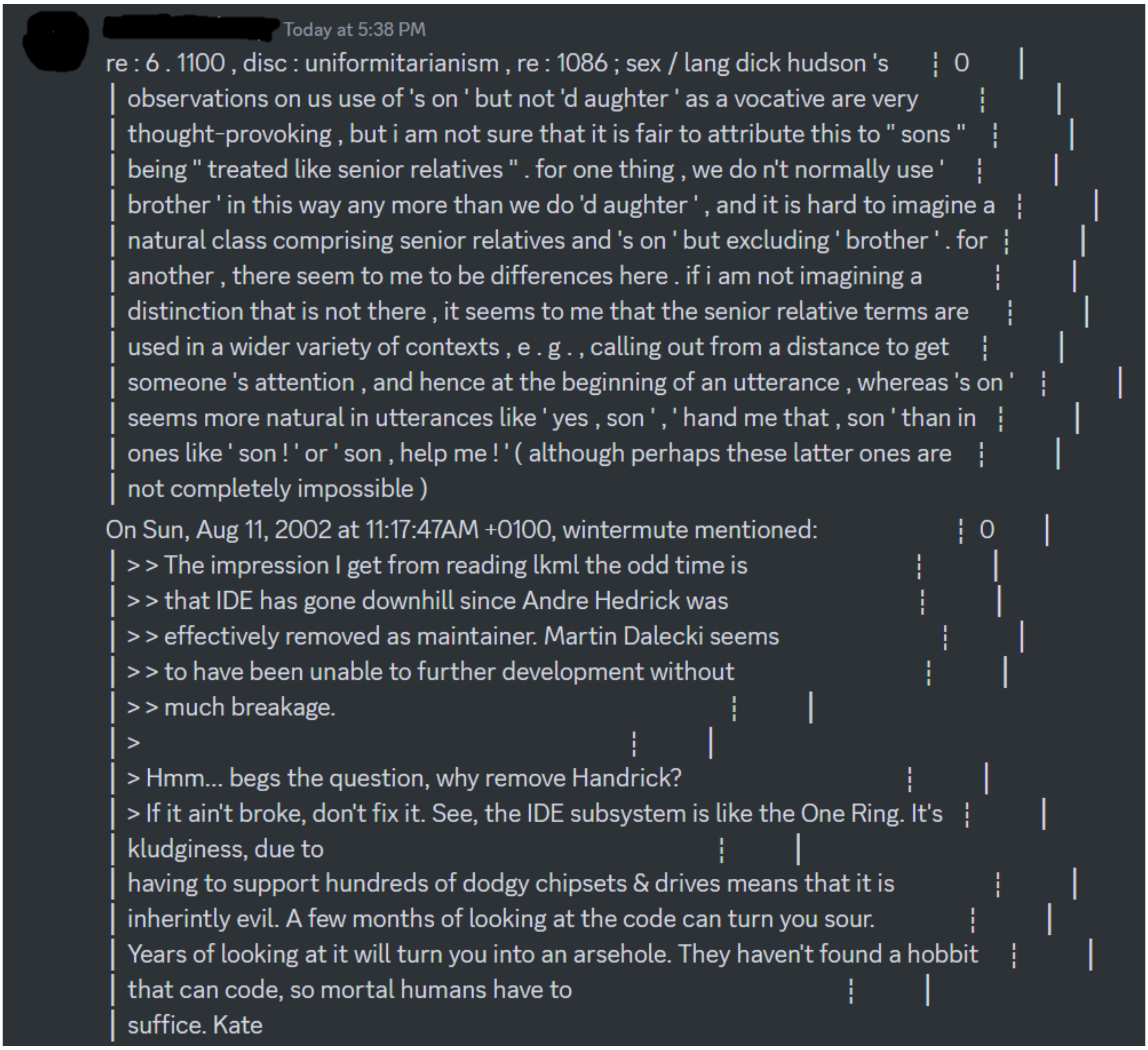
This message is considered phishing!!



```
[redacted] [redacted] Today at 5:37 PM
PROMOTE YOUR PRODUCT OR                          ¦     |
| SERVICE TO MILLIONS TODAY!                     ¦     |
| Â                                     ¦     |
| E-MAIL MARKETING                            ¦     |
| - Bulk e-mail will make you money so fast, your head will      ¦     |
| spin!- Our customers tell us that they would no longer be in    ¦     |
| business without it.- New collection of software allows you to collect targeted  ¦     |
| e-mail addresses for free!Â Â Â  See this product's web       ¦     |
| page                                     ¦     |
| CLICK                                    ¦     |
| HERE                                     ¦     |
```

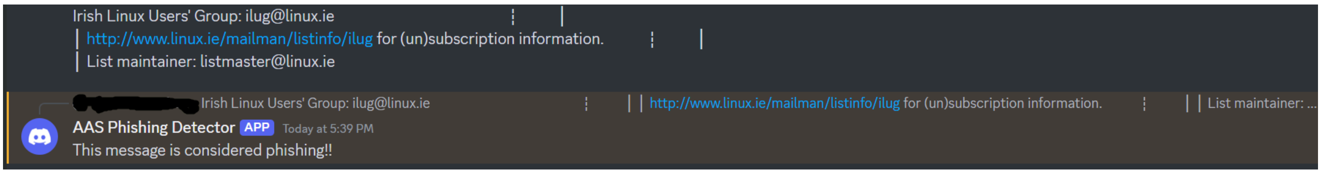
```
[redacted] PROMOTE YOUR PRODUCT OR   ¦   | | SERVICE TO MILLIONS TODAY!   ¦   | | Â
AAS Phishing Detector APP Today at 5:37 PM
This message is considered phishing!!
```

It also successfully identified benign messages (the bot doesn't reply to the message when its considered benign).



```
[redacted] [redacted] Today at 5:38 PM
re : 6 . 1100 , disc : uniformitarianism , re : 1086 ; sex / lang dick hudson 's   ¦ 0   |
| observations on us use of 's on ' but not 'd aughter ' as a vocative are very   ¦   |
| thought-provoking , but i am not sure that it is fair to attribute this to " sons "   ¦   |
| being " treated like senior relatives " . for one thing , we do n't normally use '   ¦   |
| brother ' in this way any more than we do 'd aughter ' , and it is hard to imagine a   ¦   |
| natural class comprising senior relatives and 's on ' but excluding ' brother ' . for   ¦   |
| another , there seem to me to be differences here . if i am not imagining a   ¦   |
| distinction that is not there , it seems to me that the senior relative terms are   ¦   |
| used in a wider variety of contexts , e . g . , calling out from a distance to get   ¦   |
| someone 's attention , and hence at the beginning of an utterance , whereas 's on '   ¦   |
| seems more natural in utterances like ' yes , son ' , ' hand me that , son ' than in   ¦   |
| ones like ' son ! ' or ' son , help me ! ' ( although perhaps these latter ones are   ¦   |
| not completely impossible )
On Sun, Aug 11, 2002 at 11:17:47AM +0100, wintermute mentioned:   ¦ 0   |
| > > The impression I get from reading lkml the odd time is   ¦   |
| > > that IDE has gone downhill since Andre Hedrick was   ¦   |
| > > effectively removed as maintainer. Martin Dalecki seems   ¦   |
| > > to have been unable to further development without   ¦   |
| > > much breakage.   ¦   |
| >   ¦   |
| > Hmm... begs the question, why remove Handrick?   ¦   |
| > If it ain't broke, don't fix it. See, the IDE subsystem is like the One Ring. It's   ¦   |
| kludginess, due to   ¦   |
| having to support hundreds of dodgy chipsets & drives means that it is   ¦   |
| inherintly evil. A few months of looking at the code can turn you sour.   ¦   |
| Years of looking at it will turn you into an arsehole. They haven't found a hobbit   ¦   |
| that can code, so mortal humans have to   ¦   |
| suffice. Kate
```

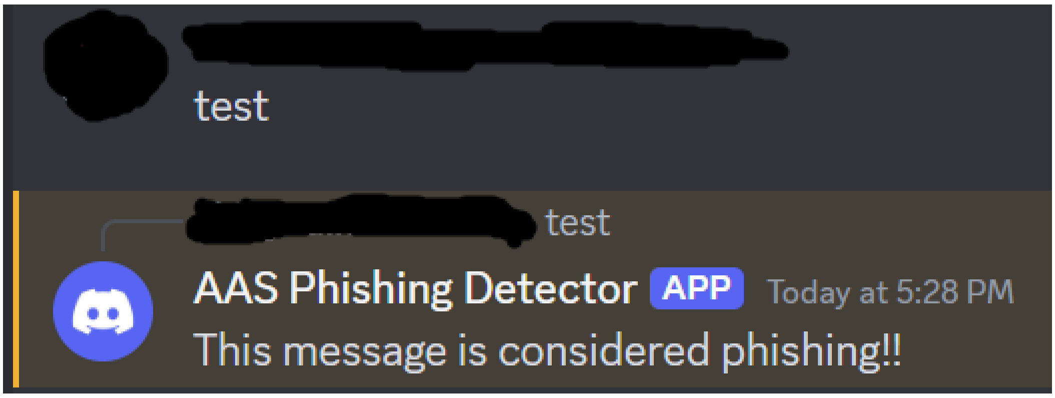
However, its not perfect. It considered some benign messages as phishing messages.



Irish Linux Users' Group: ilug@linux.ie
| http://www.linux.ie/mailman/listinfo/ilug for (un)subscription information.
| List maintainer: listmaster@linux.ie

[redacted] Irish Linux Users' Group: ilug@linux.ie | | http://www.linux.ie/mailman/listinfo/ilug for (un)subscription information. | | List maintainer: ...

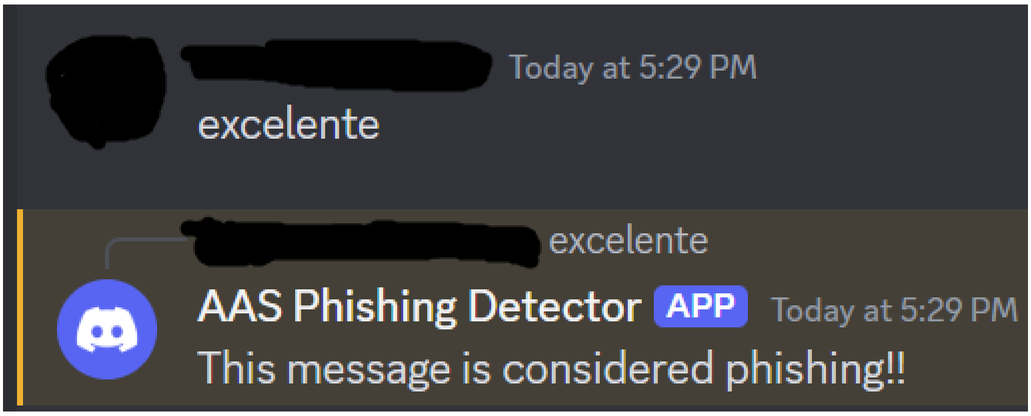
AAS Phishing Detector APP Today at 5:39 PM
This message is considered phishing!!



[redacted] [redacted]
test

[redacted] test

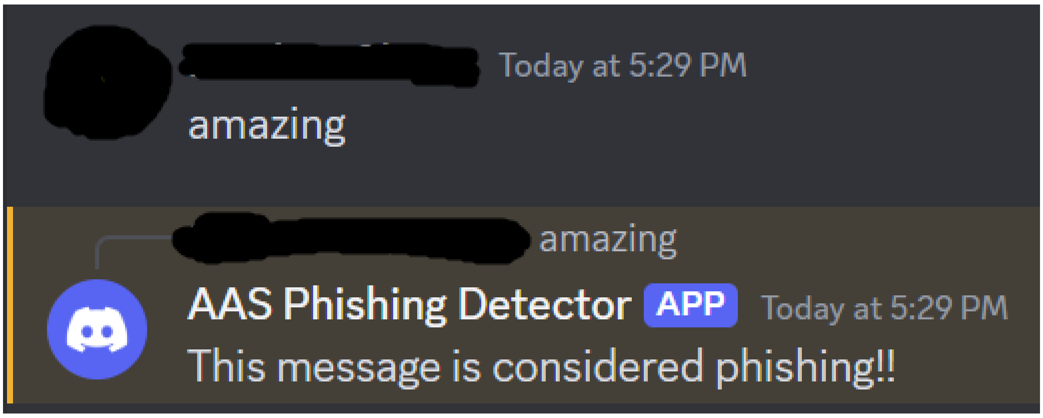
AAS Phishing Detector APP Today at 5:28 PM
This message is considered phishing!!



[redacted] [redacted] Today at 5:29 PM
excelente

[redacted] excelente

AAS Phishing Detector APP Today at 5:29 PM
This message is considered phishing!!



[redacted] [redacted] Today at 5:29 PM
amazing

[redacted] amazing

AAS Phishing Detector APP Today at 5:29 PM
This message is considered phishing!!

We suspect that its due to the `Fast Text` model not being trained well enough.

Nevertheless, the model and the bot worked as intended.
Albeit not perfect, it detects phishing messages.