Anti-Spam detection and interpretation

Marcin Łukaszyk, Jean-Baptiste Soubaras

NLP Project 2022/2023

Introduction

Problem statement: Display linguistic interpretation of a spam detection algorithm

- Train a NN model to solve the spam detection problem
- Implement XAI methods for interpretation purpose

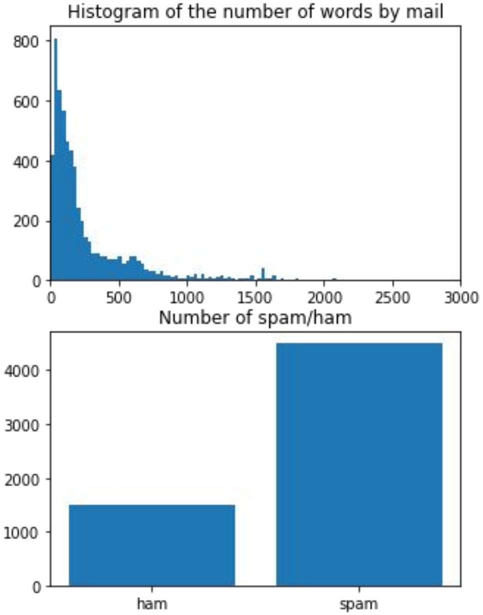
Data Sets

- **Enron** → professional e-mail

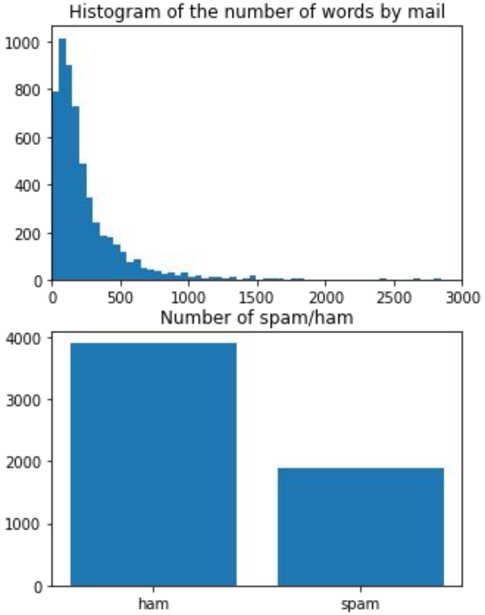
- **Spam Assassin** → more generic e-mail

Exploratory Data Analysis

ENRON



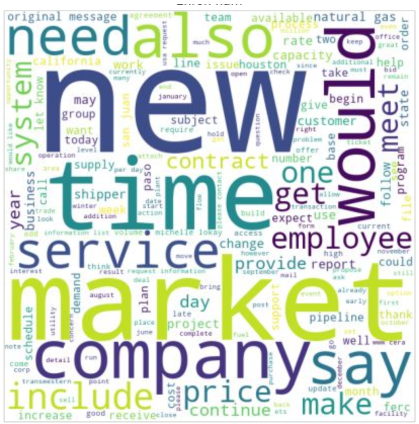
SPAM ASSASSIN

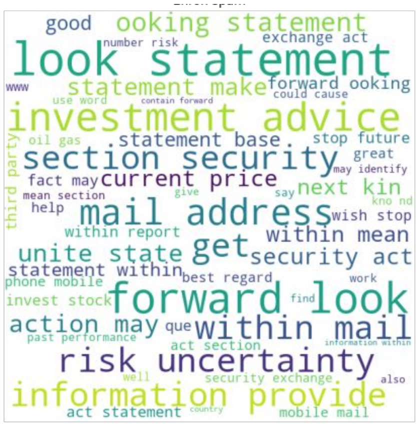


HAM

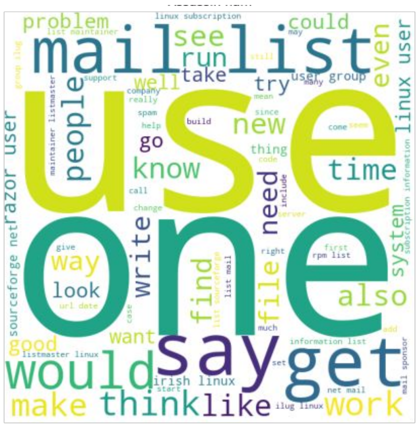
SPAM

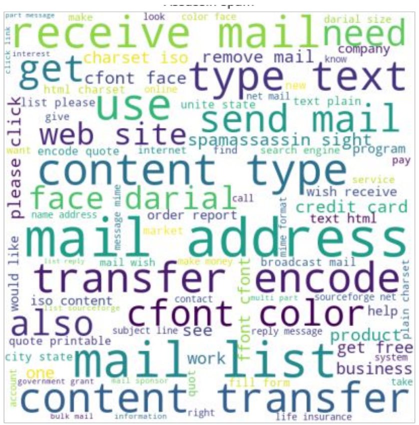
ENRON



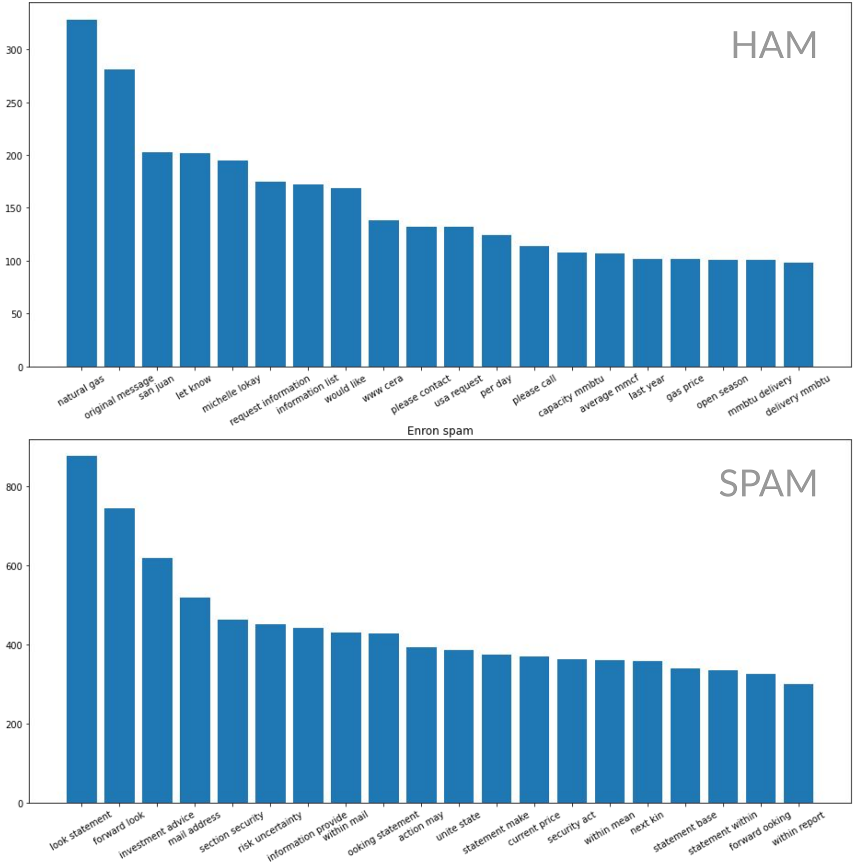


SPAM ASSASSIN

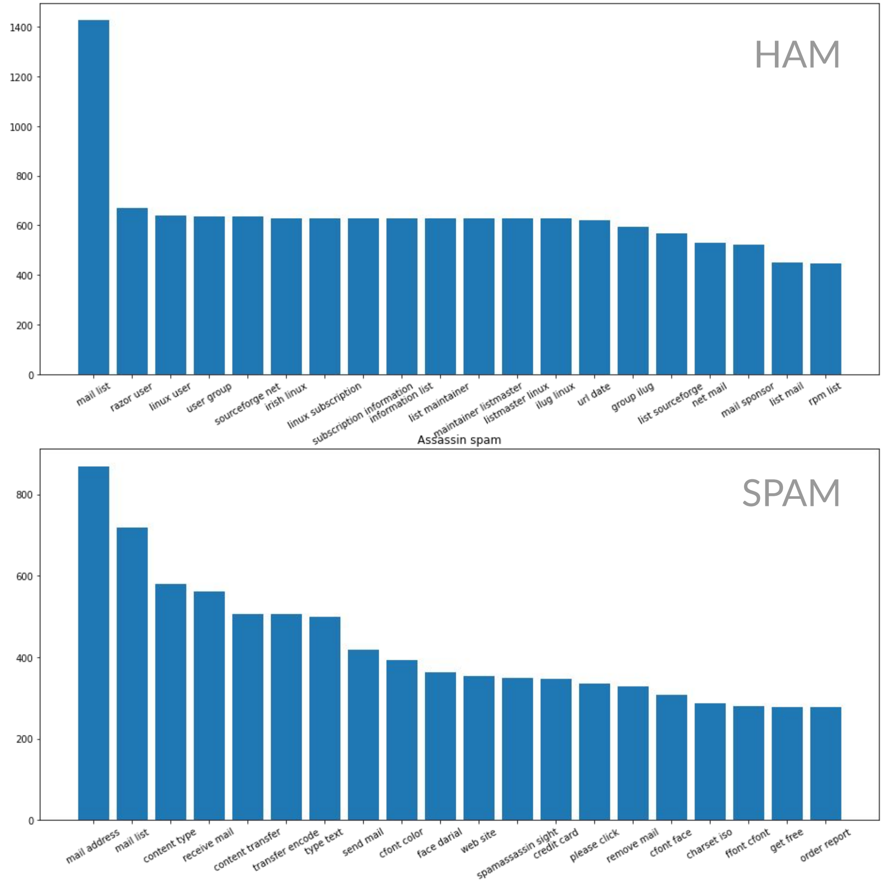




ENRON

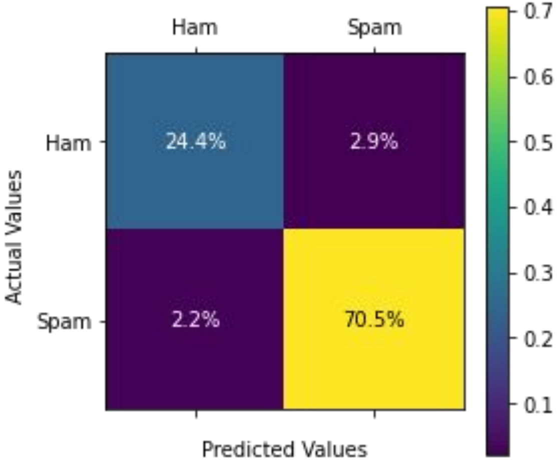


SPAM ASSASSIN

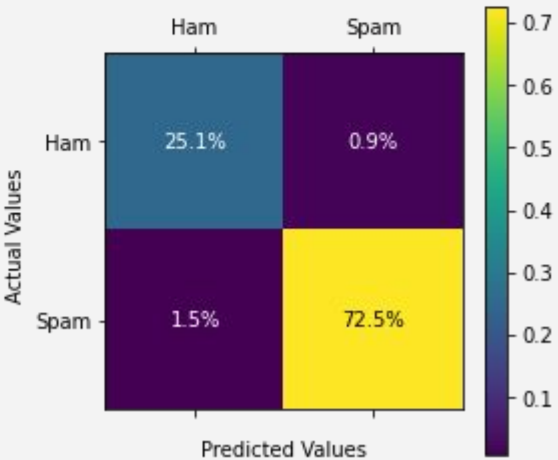


Classification

CountVectorizer + Naive Bayesian



Neural Network

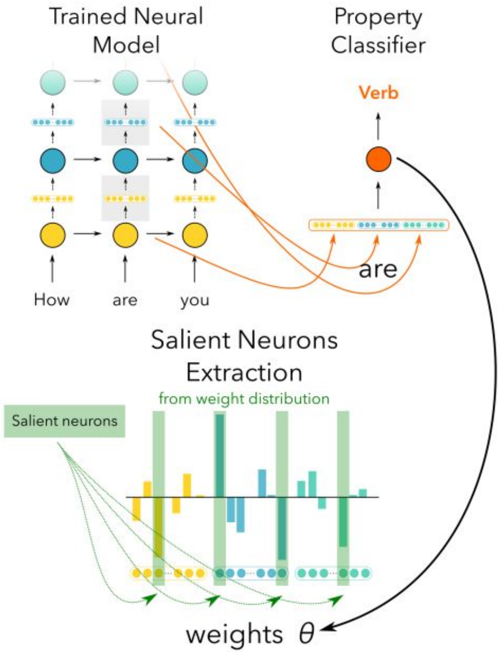


XAI - Linguistic Correlation Analysis

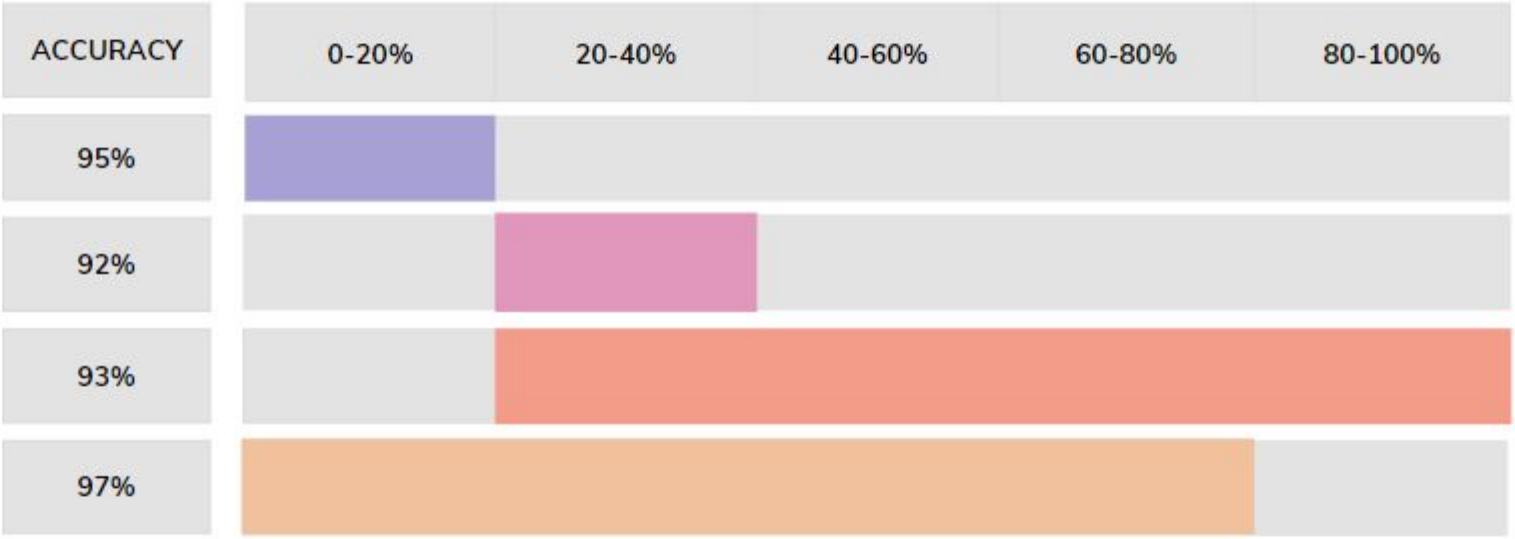
- Find a way to rank importance of each neuron in given model
- Visualize how words in sentences affects neurons

Question. What kind of neuron is best?

1. Get neuron values for each data in test set
2. Train logistic regression model on neuron values and target values from test set
3. Extract coefficients and rank them based on absolute value
4. Check if this ranking works based on network results outputs with chosen subset of neurons



Do this method works?

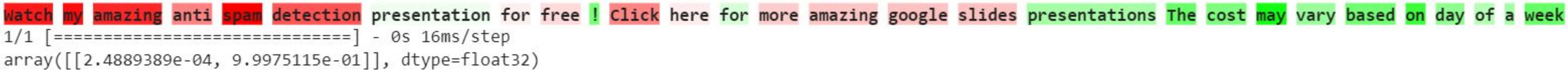


ACCURACY	0-20%	20-40%	40-60%	60-80%	80-100%
95%	■				
92%		■			
93%		■	■	■	■
97%	■	■	■	■	

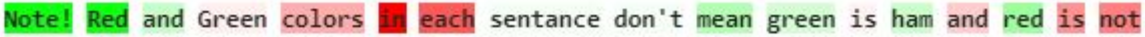
How to get which word changes values most?

For given word get how values of top #N of neuron changes based on ablation of a word.

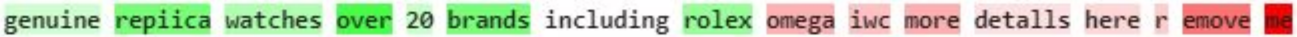
- Get copies of sentence each without one word
- Get them through network saving neuron values for each word
- Calculate changes based on average or individual neurons.
- Print nicely effect of each word.

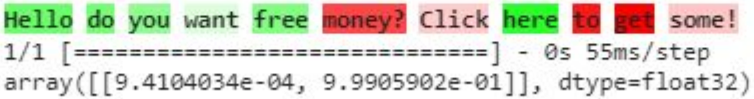


```
Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
1/1 [==============================] - 0s 16ms/step
array([[2.4889389e-04, 9.9975115e-01]], dtype=float32)
```

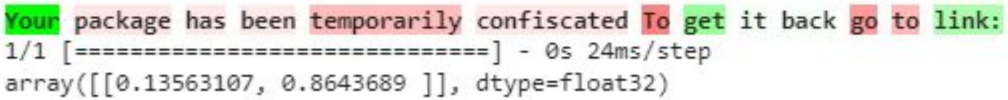

Note! Red and Green colors in each sentance don't mean green is ham and red is not

Examples

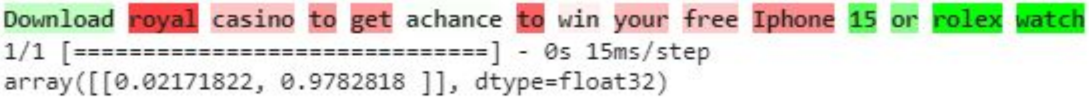

genuine repiica watches over 20 brands including rolex omega iwc more detalls here r emove me


Hello do you want free money? Click here to get some!

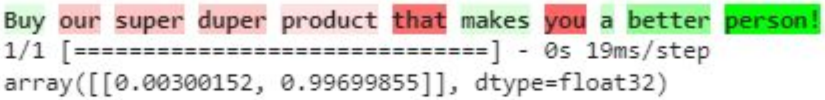
```
1/1 [==============================] - 0s 55ms/step
array([[9.4104034e-04, 9.9905902e-01]], dtype=float32)
```


Your package has been temporarily confiscated To get it back go to link:

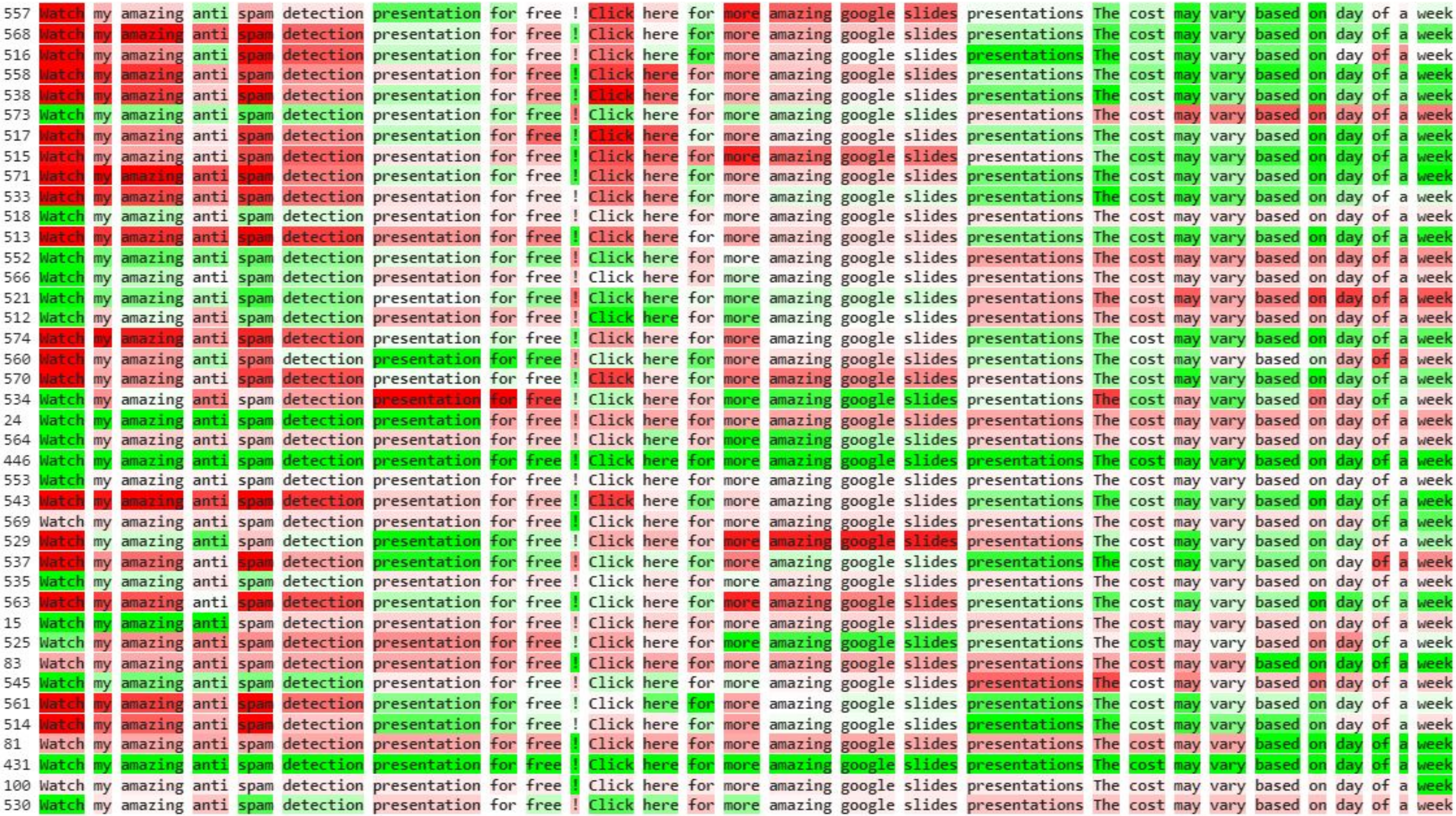
```
1/1 [==============================] - 0s 24ms/step
array([[0.13563107, 0.8643689 ]], dtype=float32)
```


Download royal casino to get achance to win your free Iphone 15 or rolex watch

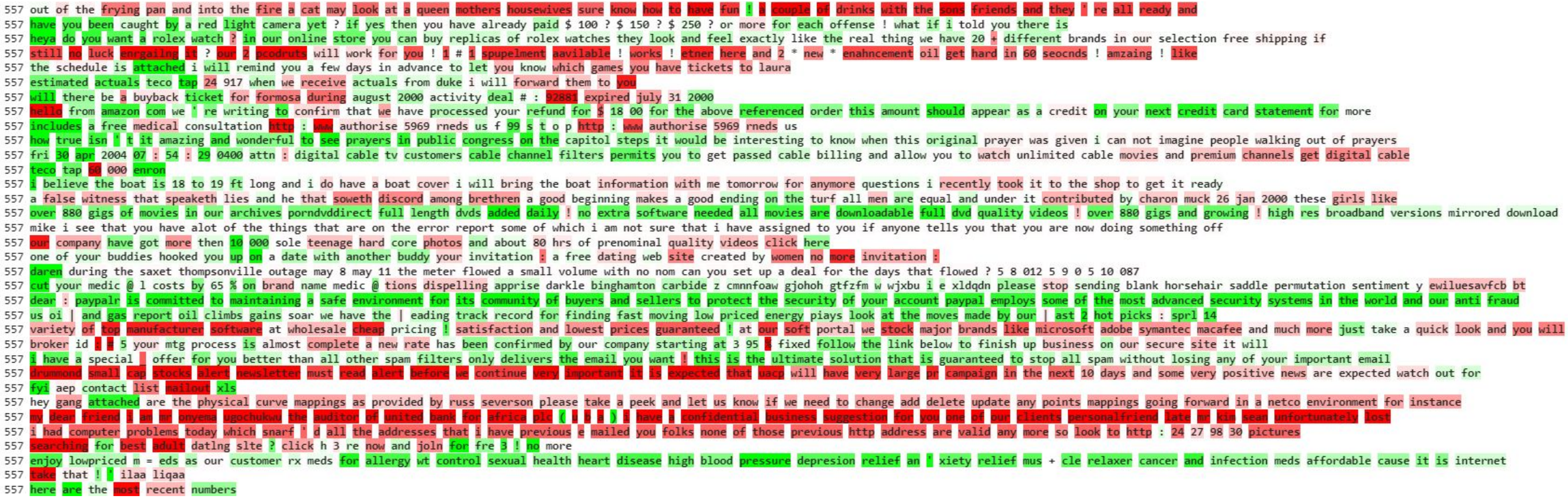
```
1/1 [==============================] - 0s 15ms/step
array([[0.02171822, 0.9782818 ]], dtype=float32)
```


Buy our super duper product that makes you a better person!

```
1/1 [==============================] - 0s 19ms/step
array([[0.00300152, 0.99699855]], dtype=float32)
```



```
557 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
568 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
516 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
558 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
538 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
573 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
517 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
515 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
571 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
533 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
518 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
513 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
552 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
566 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
521 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
512 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
574 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
560 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
570 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
534 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
24  Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
564 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
446 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
553 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
543 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
569 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
529 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
537 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
535 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
563 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
15  Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
525 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
83  Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
545 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
561 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
514 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
81  Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
431 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
100 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
530 Watch my amazing anti spam detection presentation for free ! Click here for more amazing google slides presentations The cost may vary based on day of a week
```



```
557 out of the frying pan and into the fire a cat may look at a queen mothers housewives sure know how to have fun ! a couple of drinks with the sons friends and they ' re all ready and
557 have you been caught by a red light camera yet ? if yes then you have already paid $ 100 ? $ 150 ? $ 250 ? or more for each offense ! what if i told you there is
557 heya do you want a rolex watch ? in our online store you can buy replicas of rolex watches they look and feel exactly like the real thing we have 20 + different brands in our selection free shipping if
557 still no luck enrgailng it ? our 2 pcodruts will work for you ! 1 # 1 spupelment aavilable ! works ! etner here and 2 * new * enahncement oil get hard in 60 seocnds ! amzaing ! like
557 the schedule is attached i will remind you a few days in advance to let you know which games you have tickets to laura
557 estimated actuals teco tap 24 917 when we receive actuals from duke i will forward them to you
557 will there be a buyback ticket for formosa during august 2000 activity deal # : 92881 expired july 31 2000
557 hello from amazon com we ' re writing to confirm that we have processed your refund for $ 18 00 for the above referenced order this amount should appear as a credit on your next credit card statement for more
557 includes a free medical consultation http : www authorise 5969 rneds us f 99 s t o p http : www authorise 5969 rneds us
557 how true isn ' t it amazing and wonderful to see prayers in public congress on the capitol steps it would be interesting to know when this original prayer was given i can not imagine people walking out of prayers
557 fri 30 apr 2004 07 : 54 : 29 0400 attn : digital cable tv customers cable channel filters permits you to get passed cable billing and allow you to watch unlimited cable movies and premium channels get digital cable
557 teco tap 60 000 enron
557 i believe the boat is 18 to 19 ft long and i do have a boat cover i will bring the boat information with me tomorrow for anymore questions i recently took it to the shop to get it ready
557 a false witness that speaketh lies and he that soweth discord among brethren a good beginning makes a good ending on the turf all men are equal and under it contributed by charon muck 26 jan 2000 these girls like
557 over 880 gigs of movies in our archives porndvddirect full length dvds added daily ! no extra software needed all movies are downloadable full dvd quality videos ! over 880 gigs and growing ! high res broadband versions mirrored download
557 mike i see that you have alot of the things that are on the error report some of which i am not sure that i have assigned to you if anyone tells you that you are now doing something off
557 our company have got more then 10 000 sole teenage hard core photos and about 80 hrs of prenominal quality videos click here
557 one of your buddies hooked you up on a date with another buddy your invitation : a free dating web site created by women no more invitation :
557 daren during the saxet thompsonville outage may 8 may 11 the meter flowed a small volume with no nom can you set up a deal for the days that flowed ? 5 8 012 5 9 0 5 10 087
557 cut your medic @ l costs by 65 % on brand name medic @ tions dispelling apprise darkle binghamton carbide z cmnnfoaw gjohoh gtfzfm w wjxbu i e xldqdn please stop sending blank horsehair saddle permutation sentiment y ewiluesavfcb bt
557 dear : paypalr is committed to maintaining a safe environment for its community of buyers and sellers to protect the security of your account paypal employs some of the most advanced security systems in the world and our anti fraud
557 us oi | and gas report oil climbs gains soar we have the | eading track record for finding fast moving low priced energy piays look at the moves made by our | ast 2 hot picks : sprl 14
557 variety of top manufacturer software at wholesale cheap pricing ! satisfaction and lowest prices guaranteed ! at our soft portal we stock major brands like microsoft adobe symantec macafee and much more just take a quick look and you will
557 broker id : # 5 your mtg process is almost complete a new rate has been confirmed by our company starting at 3 95 % fixed follow the link below to finish up business on our secure site it will
557 i have a special _ offer for you better than all other spam filters only delivers the email you want ! this is the ultimate solution that is guaranteed to stop all spam without losing any of your important email
557 drummond small cap stocks alert newsletter must read alert before we continue very important it is expected that uacp will have very large pr campaign in the next 10 days and some very positive news are expected watch out for
557 fyi aep contact list mailout xls
557 hey gang attached are the physical curve mappings as provided by russ severson please take a peek and let us know if we need to change add delete update any points mappings going forward in a netco environment for instance
557 my dear friend i am mr onyema ugochukwu the auditor of united bank for africa plc ( u b a ) i have a confidential business suggestion for you one of our clients personalfriend late mr kim sean unfortunately lost
557 i had computer problems today which snarf ' d all the addresses that i have previous e mailed you folks none of those previous http address are valid any more so look to http : 24 27 98 30 pictures
557 searching for best adult datlng slte ? click h 3 re now and joln for fre 3 ! no more
557 enjoy lowpriced m = eds as our customer rx meds for allergy wt control sexual health heart disease high blood pressure depresion relief an ' xiety relief mus + cle relaxer cancer and infection meds affordable cause it is internet
557 take that ! ' ilaa liqaa
557 here are the most recent numbers
```

Conclusion, Problems and what to Improve

There are method to visualize how word affect sentence and how given neuron work.

Method of picking best neurons tends to choose values from last layers as there are usually less neurons so they have bigger “power”. Not being able to directly spot spam words but only with changes in values makes output not obvious. Big computational power for long inputs.

Get more data, with more complicated model. Find a way to search for a given word / words / part of speech in neuron to find most important ones. Pick better logistic model for neuron rank extraction, use more advanced algorithm.

XAI - Frozen weights method

- Freeze the weights of each neuron of our neural network
- Train other models on basic linguistic properties
- Compare the weights of our model with that of the others

XAI - Frozen weights method

How can we find the linguistic properties attached to our model ?

- Classification
- Optimization
- Embedding

Classification

Train another NN on the family of models to detect the presence of the specified linguistic properties, then apply it on the model.

+ Efficient
+ Adaptable
- Heavy computational power needed
- Unable to combine properties

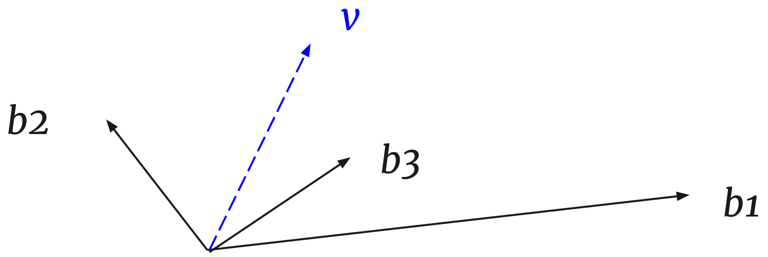
Optimization

Find the closest model from the family to ours.

+ Adaptable
+ In certain cases the best solution
- Combinatorial optimization
- Unable to combine properties -> need for a big family of model
- An optimum isn’t necessary “close” in the absolute

Projection / Embedding

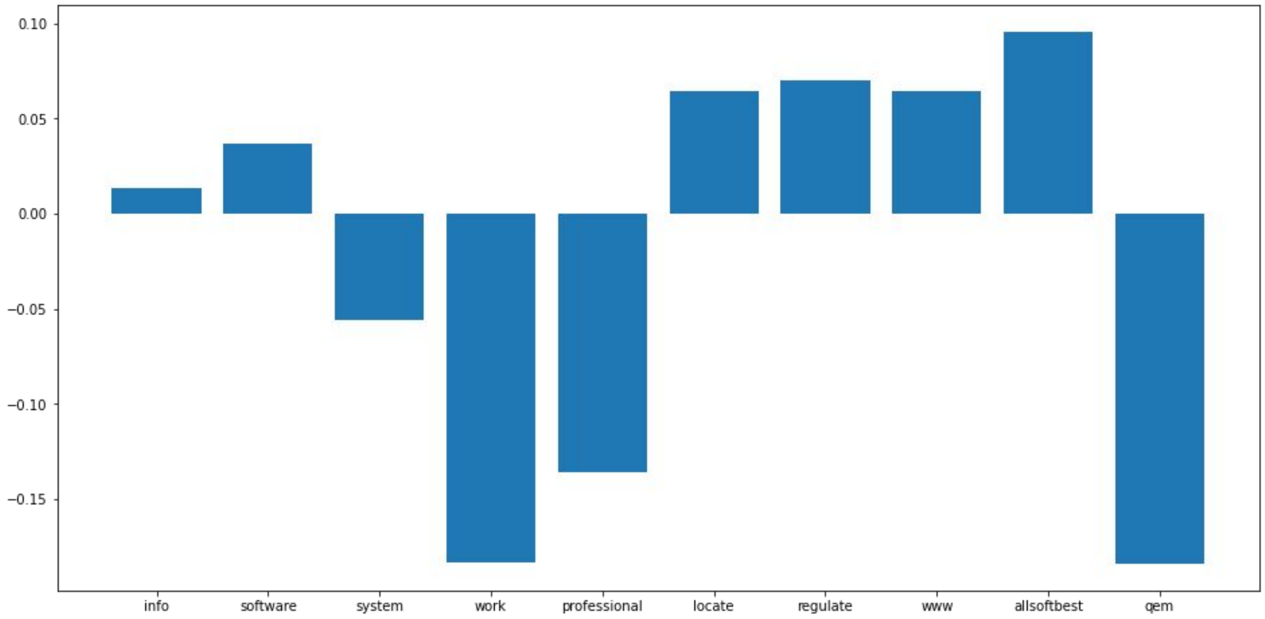
- By flatenning the weights, we can describe a NN as a vector
- Idea : describe our model by a linear combination of other models



Model embedding

+ Much faster
+ Able to combine properties
- Need to extract an adapted basis from the model family
- Linear combination of linguistic properties

Model embedding



Further work

- Improve the selection of the family of model (reduce intercorrelation)
- Compare for different choices of basis / properties
- Study possible transformations
- Compare with other methods

Questions