Anti-Spam : A new methodology for interpreting anti-spam classification
Project Proposal for NLP Course, Winter 2022

Marcin Łukaszyk
WUT
Student
01133055@pw.edu.pl

Jean-Baptiste Soubaras
WUT
Student
01182889@pw.edu.pl

Supervisor: Anna Wróblewska
WUT
anna.wroblewska1@pw.edu.pl

Abstract

The war on spam e-mails has been an important issue in the recent decades, and the development of Machine Learning (ML) methods - especially that of Recurrent Neural Networks (RNN) - in the Natural Language Processing (NLP) domain led to always more accurate AI-based filters able to classify e-mails are *spam* or *ham*. In return, the behavior of these algorithms has been more and more opaque for human understanding, giving very few insights on the resolution of the problem as well as on the limitations of the algorithms. The subject of the project described in this proposal is to research Explainable Artificial Intelligence (XAI) methods to solve the spam detection problem while giving humanly-interpretable insights on the deeper functioning of the ML algorithm. The global purpose would be to elaborate a methodology to make any AI-based classifier more interpretable. To proceed, the project will be divided in two parts : the first aiming at the writing of an efficient AI-based spam filtering algorithm, the second at the exploitation of this algorithm using diverse XAI methods in order to make it more explanatory.

1 Introduction

1.1 Background

The outburst of the Internet in the late 90s was a phenomenon that revolutionised the way of communicating. In comparison with the traditional post mail, the arrival of the electronic mail, more commonly named e-mail, and its wide-spreading use allowed for much quicker and more convenient message delivery. However, such a convenient media would quickly show a few drawbacks. One of them is the overabundance of spam, also called junk mails.

A spam is an irrelevant or unsolicited message sent by mail, generally to a large number of users, for the purposes of advertising, phishing (i.e. tricking someone into giving access to their credentials or bank data for malicious purpose), spreading malware, etc. It is estimated that the proportion of spam among the global em-mail traffic is about 50%. It has thus been an issue for mailbox providers to ensure that their users would not be exposed to such content, by filtering the incoming messages and programmatically spotting the ones suspected to be spam, generally to redirect them to another box labeled as ”SPAM”. Most of the algorithms assigned to this task use Machine Learning (ML) methods to achieve their goal.

The main problem caused by ML algorithms is that they act like a black box, able to classify a given input but without being able to give humanly-interpretable reasons for their choice. As a result, the algorithm is unable to provide an understanding on the problem, in spite of its ingestion of large sets of data. In the recent years, many researchers have worked on the interpretability of such algorithms, in order to find methods for using the ability to train on a large number of data as a way of obtaining a better understanding on a problem.

The problem of spam detection is quite adapted for tackling this issue for two reasons : first, it is a problem that modern ML algorithms tackle quite easily (for instance, *Google* and its mailbox *Gmail* claim an accuracy of 99.9% of their spam and phishing detection algorithms); second, human analysis and understanding already give a lot of insights on the characteristics that make an e-mail a spam e-mail. Actually, a lot of spam senders display some hints in their message (spelling mistakes, unrealistic promises,...) in order to target the less vigilant people that will follow their instructions until the end and are less likely to engage in judiciary process). It is thus a good topic for implementing interpretation methods for ML algorithms and analysing their outputs.

1.2 Description of the subject

The main goal of the project is to analyse and explain predictions of machine learning models resolving the spam detection problem. Knowing how different inputs affect outputs and why output changes helps build trust and brings AI closer to general audience. Moreover XAI (Explainable artificial intelligence) can be used by researchers to confirm existing domain knowledge or to discover new insights. It can also be used to look for any existing bias in machine learning models.

1.3 Significance of the project

Applying deep learning techniques to NLP problems had greatly improved scores and ability to solve complex task. Contrarily it had negatively affected how attainable is to interpreted inner workings of complex machine learning models. To counter lost information it is necessary to ellaborate methods of making models more explainable and interpretable. During our work we will try to use techniques like: feature importance, surrogate model and visualizations of model predictions.

Spam and spam detection started as soon as internet become something publicly available in mid 1990 although first spam email is dating back to May 3, 1978. First methods of dealing with spam were basically blocking given addresses. Mainly in form of black hole lists with blacklisted addresses. Other methods were based on DNS addresses. During 2000 spam became serious problem as it impacted user experience and many computer viruses were transformed using spam emails. To fight it many organizations (public and private) created various algorithms and legal measures. They used techniques like ham passwords, Checksum-based filtering or first NLP methods based on regular expressions.

The global improvements in spam filtering algorithm have led to a decline in the total amount of spam sent each year.

Interpretability and explanability is vital for further as making sure how and why given algorithms work help ensure credibility of reserchers. To keep up with new, more advanced methods and algorithms for NLP researchers must create and improve various methods to explain models behavior.

1.4 Plan of the report

2 Related Work

2.0.1 Current algorithms

Most of machine learning algorithms can be applied to task of spam detection. Most popular ”standard” ones are KNN, Naïve Bayes and Reverse DBSCAN (Harisinghaney et al., 2014) or random forest. More advanced models using deep learing are also in use like standard deep learning neural networks, Long Short Term Memory networks (AbdulNabi and Yaseen, 2021) or Convolutional Neural Network and Multi-Layer Perceptron (Shahariar et al., 2019).

2.1 Future perspectives for neural networks interpretability

Interpretability and explanability is vital for further as making sure how and why given algorithms work help ensure credibility of reserchers. To keep up with new, more advanced methods and algorithms for NLP researchers must create and improve various methods to explain models behavior.

2.1.1 Methods employed

(Belinkov and Glass, 2019) provides a benchmark on several methods used for interpreting classifiers based on neural networks. Many methods rely on the prediction of linguistic properties from activation of the neural network. In this approach a first neural network model is trained on the main problem. Then, the trained model is used for generating feature representations. Finally another classifier is used to predict the linguistic property of interest.

Another approach suggests linguistic Correlation Analysis and Cross-model Correlation Analysis with comprehensive analysis of neurons to analyse distribution of different linguistic properties and neurons exclusivity to some properties.

First method, linguistic Correlation Analysis, trains second, some kind of linear model for easy explanability, based on neuron activations values from first, already trained on our dataset model, with labels from our original dataset. Then based on absolute values of weights in new, second model we can deduce ranking of neurons importance of first model, trained to detect spam.

Cross-model Correlation Analysis works by training multiple similar (”using identical model settings but with differing training data and initialization”) models and then for each neuron in our architecture compering Pearson correlation coefficient between neuron activation values in original models and ones without neurons. Then we can rank them based on correlation coefficient and deduce most important ones.

2.2 Data Sets

A preliminary research allowed to find several available data sets.

SpamAssasin - 2001 / 6047 instances
This data set has been created by the *Apache Software Fondation* to develop their spam filter software.

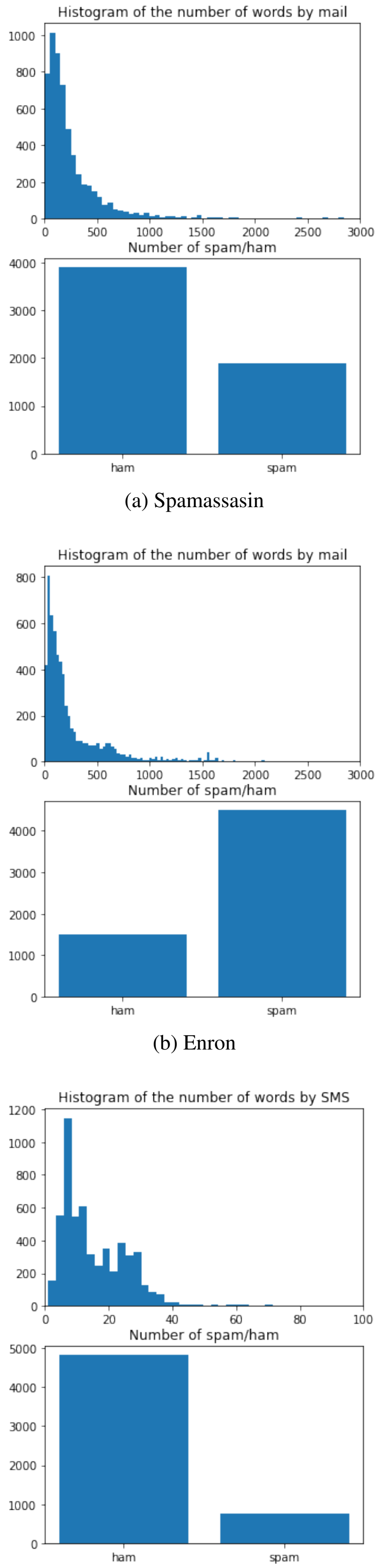
Enron-Spam - 2006 / 6000 instances
The data set was created and studied in the following article (V. Metsis and Paliouras, 2006).

SMS Spam Collection - 2012 / 5574 instances
The data set was created and studied in the following article (Almeida, 2011). It contains SMS and not e-mails, but it will be used to compare the interpretations obtained on different formats of text.

Two additional data sets dedicated to this end were also found, however for privacy reasons these data sets were already encoded and are not suitable for language interpretation. In consequence, they will only be considered for testing purposes if needed.

Spambase dataset - 1999 / 4601 instances
The data set was created by (Mark Hopkins, 1999). The file contains a classification based on 57 features of e-mails, most of them being the frequency of a word or character. There is no acces to the raw text data.

PU Corpora - 2003
The data set was created by (I. Androutsopoulos, 2003). The e-mails contained in the data set are all tokenized for privacy matters.



(a) Spamassasin

(b) Enron

(c) SMS Spam Collection

Figure 1: Content of the data sets.

3 Approach and Methodology

3.1 Functional methodology

1. Preliminary bibliographical research
2. Creation of a work environment
3. Implementation of simple spam detection models
4. Evaluation and analysis of the spam detection
5. Implementation of known methods for interpretation of the model
6. Evaluation and analysis
7. Research and improvement -¿ added value of our work
8. Evaluation and analysis
9. Report and presentation

3.2 Evaluation and analysis

The different methods that will be studied have been detailed in section 3. Considering the primary spam detection data set, that latter will be evaluated on a test data set. The purpose is to use neural networks to train an efficient enough model, yet opaque in its execution.

For XAI algorithms, evaluation will be made by comparison of the interpretation computed with human interpretation and comparison of the output with that of the primary algorithm. The project will use an additional data set consisting of SMS spam to compare the differences of interpretation between e-mail spam and SMS spam. Eventually, if relevant enough, an implementation of naive heuristic methods based on the insights given by the XAI methods could be compared with the primary algorithm.

3.3 Tools

The tools that will be used will be detailed in this part.

Hardware

- Personal computers;
- In case of a need for heavy computation (training ML models on large data sets), the use of the computers of the laboratories of WUT could be envisaged.

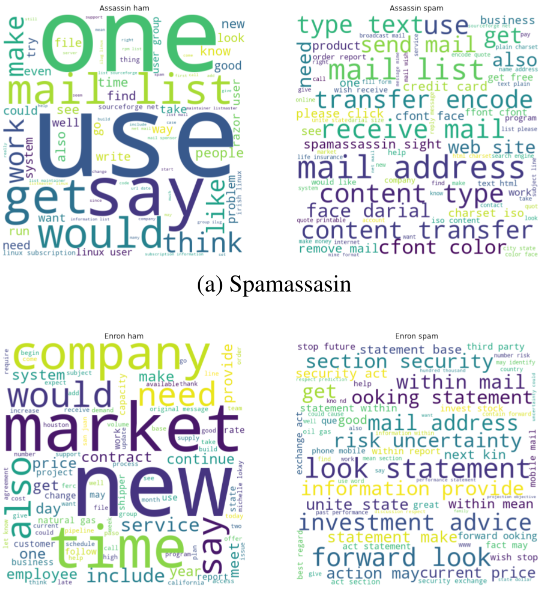
Software

- Programation language: Python 3.6;
- Libraries: Natural Langueage Toolkit, Keras, Tensorflow;
- Collaboration: Google Collab, maybe GitHub;
- Data sets: Kaggle connector;

4 Preliminary work and classification

4.1 Exploratory Data Analysis

For the Exploratory Data Analysis (EDA), we focused on the two main data sets : *Enron* and *SpamAssassin*. The data sets were first parsed and preprocessed to remove all punctuation, digit and stop word, and to lemmatize all the remaining words. Then, in Figure 2, the word cloud of each data set was plotted according to the class of each mail : ”spam” or ”ham”.



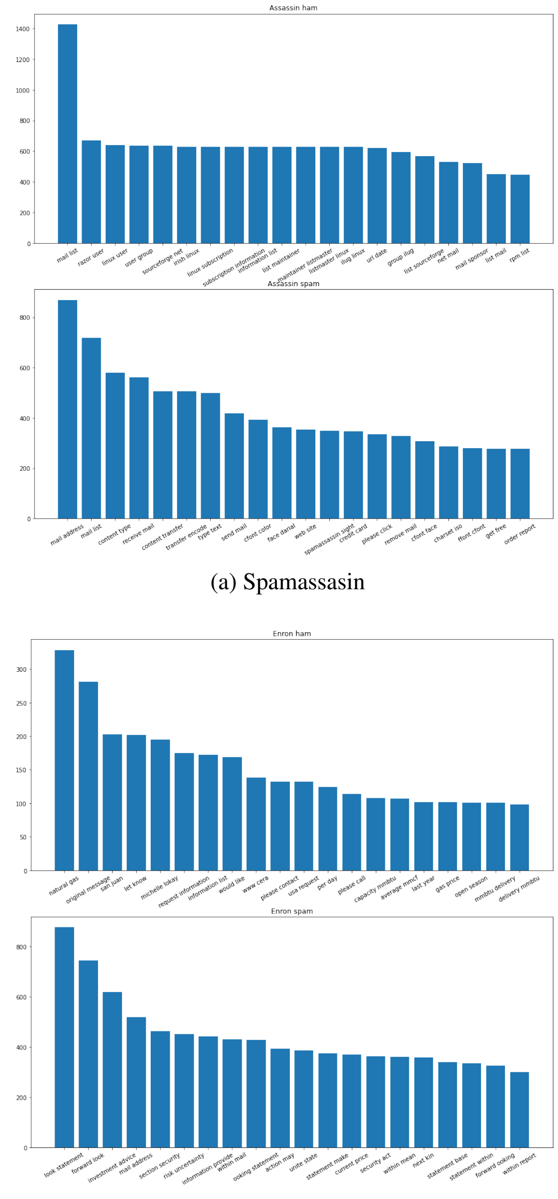
(a) Spamassasin

(b) Enron

Figure 2: Word clouds of the data sets.

We can observe first that the Enron data sets uses much more words coming from the professional language (”market”, ”company”, ”employee”,...) whereas in SpamAssassin the most common words are words of the everyday language. Moreover, in both cases the spam mails seem to use less of these overly employed words, but instead contains a lot of occurrences of expressions that would in a lot of cases seem suspicious (for instance, very few ”investment advices” are given by mail).

To get more first hand insights, the repartition of the most common bigrams were also plotted in Figure 3.



(a) Spamassasin

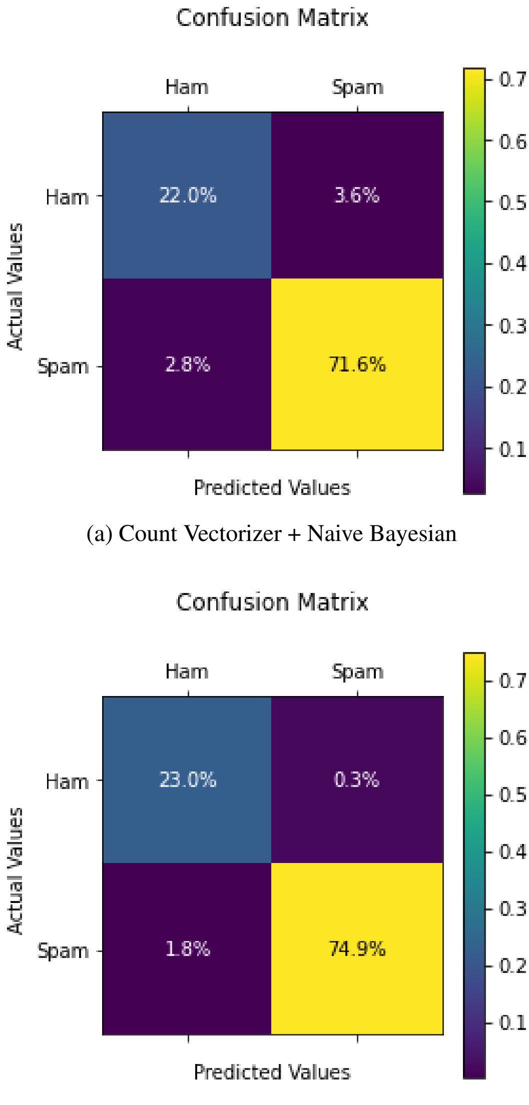
(b) Enron

Figure 3: Most common bigrams.

Again, the vocabulary is much more professional in the Enron data set, and the choice of words differ quite a lot between spam and ham mails.

4.2 Classification

The task of classification not being the heart of our research work, we implemented a generic neural network model (the sequential model from the library keras). For the sake of comparison and in to attest the quality of the model, we also implemented a Naive Bayesian algorithm on CountVectorizer embedding, which is a classical algorithm for solving our problem explained in (Christopher D. Manning, 2008). Figure 4 relates the performances of both classifier. Our Neural Network (NN) is slightly more efficient than the Naive Bayesian classifier, with a score of 98% accuracy.



(a) Count Vectorizer + Naive Bayesian

(b) Neural Network

Figure 4: Confusion matrix of each classifier.

The model we trained is thus suitable to be analysed using different XAI methods.

5 XAI methods and results

5.1 Linguistic Correlation Analysis

The linguistic correlation analysis method is used to extract most important neurons in a given task. To do that we need to find a method of ranking all neurons and then to visualize it.

Ranking all neurons is done by training logistic regression classifier on pairs $\{z_i, I_i\}$ where z_i is value of each neuron that our model use for given input x_i from original dataset and I_i is corresponding label for input in dataset. In our work we used neuron values and labels that were in testing set as to prevent any over-fit and/or data leakage. After having logistic regression model we based each neuron n_j value on absolute value of corresponding coefficient β_j.

Testing if this method works is done by observing how performance of original model changes when we will be only using some neurons. For example we can assume that this method of ranking works when model using only top 50% of neurons outperforms model where we use bottom 50% of neurons.

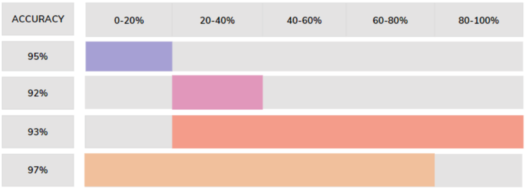
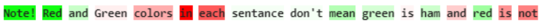


Figure 5: Model accuracy when using selected number of neurons

This proves that using logistic regression classifier to rank neurons works and sets general ranking of neuron importance.

Basic visualizations are made by calculating how lack of each word in a sentence affects neuron values. To do that from one input which is sequence of words $\{w_1, w_2, w_3...w_n\}$ we create set of n (where n is number of words in sequence) sequences where for each word there is a corresponding sequence **without** it in it. So we got set of $A = [\{w_2, w_3...w_n\}, \{w_1, w_3...w_n\}, \{w_1, w_2...w_n\}$ ... $\{w_1, w_2, w_3...w_{n-1}\}$

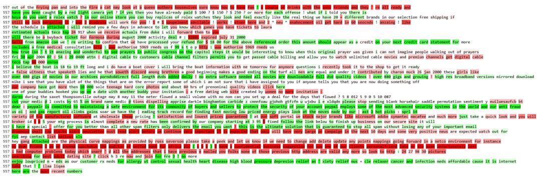
Then we predict each sequence in set using neural network model and observing how picked neurons value changes based on missing words. After getting values for each neuron and each word we can visualize it based change of neuron values.



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

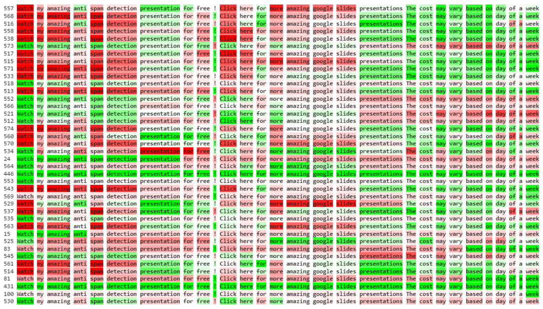
Figure 6: Example of visualization for top 100 neurons on test set example

Advanced visualizations are made by picking only one neuron and seeing how it reacts to many different sentences or by analyzing one sentence and showing changes of values for each neuron individuality. Then we can show them next to each other and compere results.



```
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 enrgaiing 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 kin 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
```

Figure 7: One neuron shown against many different sentence



```
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
```

Figure 8: Many neurons show against one sentence

5.2 Frozen Weights Method

The frozen weights method consists in freezing the inputs and outputs weights of each neuron of a trained model (see Figure 9) in order to predict the linguistic properties they describe (occurrence of a word, part of speech, meaning-based properties...). Note that the activation biases are not considered here.

To predict linguistic properties from the weights of a neural network, we need a benchmark of neural networks of the same architecture dedicated each to the detection of a specific linguistic property. In this work, we used a family of neural networks trained on the detection of one specific word each, these words being part of the vocabulary of the train data set. Another family of properties we could have used would have been a classification of the mails according to their meaning or sentiment, but it would have been harder to train the models.

To compare the weights of the main model with those of the benchmark models, we considered three possibilities:

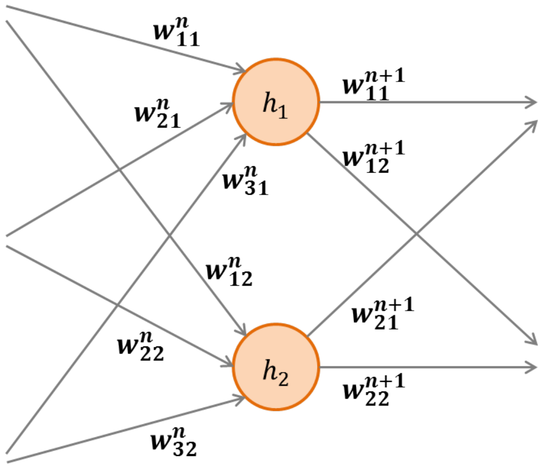


Figure 9: Weights extracted from the neurons.

- Classification;
- Optimization;
- Model embedding.

Classification is the most common approach in the literature (see (**?**)). It consists in training an external classifier (usually deep learning-based) on the benchmark model weights to predict the most accurate linguistic property that relates to the main model. This approaches has shown interesting results in the literature, however it is quite intricate and very heavy in terms of computational costs.

Optimization consists in defining a distance (by default the Euclidean distance) between the weights of two models (with respect to the used architecture). Then, the problem relates to a combinatorial optimization problem aiming at finding one or several models of the benchmark being the closest to the studied model. This allows to find the properties whose detection is a problem similar to the main spam detection problem. The drawback of this method is that combinatorial optimization is often quite intricate and usually needs a heavy use of heuristics.

Model embedding is an approach inspired by the vectorization of word embedding. It consists in flattening all the weights of the given model into a vector of an Euclidean space. Under that form, the corresponding vector can be expressed as a linear combination of the vectors of the embedded models of the benchmark. We can then project the vector into the subspace engendered by a few of the most relevant models. More formally, if M is the studied model, $B_{1:k}$ the benchmark models, $v(\cdot)$ the vectorization operation, and m the number of properties we want to observe:

1. Compute $\{|\langle v(M), v(B_i)\rangle|, \forall 0 \leq i \leq k\}$;
2. Find $i_1, ..., i_m$ the m biggest element of that set;
3. $v(M) \approx \sum_{l=1}^{m} \langle v(M), v(B_{i_l})\rangle v(B_{i_l})$.

If we consider the coefficients associated with each benchmark model as a correlation, we can express the resolution of a problem by a neural network by a combination of the detection of several characteristic linguistic properties.

In this project, the approach adopted was the latter. To obtain a family of linguistic properties relevant enough we trained for one of the mails of the test set randomly chosen one model for each word contained, each model aiming at detecting the presence of the corresponding word in any mail. These models were trained on the same train set as the original problem; even though it would have been better to train them on separate data sets to avoid biases in the data it is rather intricate to simultaneously train a large number of models each on their own data set. We then applied the model embedding method to look for the 10 most relevant words. The results are displayed Figure 10.

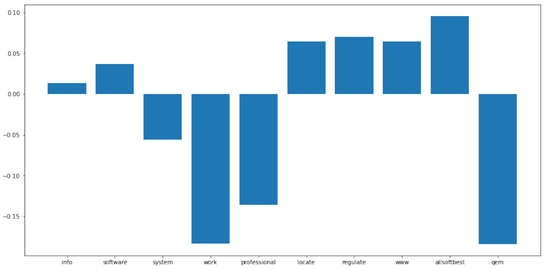


Figure 10: Results of the embedded frozen weights method.

We can observe for instance that the model is strongly negatively correlated with the detection of the word ”work”, which confirm the intuition that an e-mail containing this word is more likely to be a professional mail and has few chances to be a spam. Moreover, the word ”www” is positively correlated, which indicates that a mail containing links is more likely to be a spam. Overall, the negatively correlated words have a stronger correlation than the positively correlated, which echoes the wordcloud plotted in the EDA that showed that ham mails contained a lot of ”overused” words compared to spam mails, thus implying that the presence of these words correlates strongly with classifying the mail as ham.

5.3 Improvements to be done

Due to the short span of time dedicated to this project, comparison and evaluation of the methods haven’t had the chance to be as thorough as they were meant to be. To dig further into that direction, one should first try different parameters and variants of the methods described in this report and execute them on diverse model configurations and architecture, to observe the dependence to the structure of the neural network. Moreover, these methods should be compared to other XAI approaches applied on the spam detection problem. Unfortunately, no related work tackling that specific issue has been found for the writing of this report.

Concerning the frozen weights method, a special attention should be brought upon the choice and the training of all the benchmark model, as well as mathematical analysis of the vectorized benchmark model. One should ideally find an orthogonal basis of the space engendered by the weights of our model architecture, which surely implies applying transformation to the vectorized models. Finally, it would probably be very beneficial to try different word embeddings, vectorization for instance would surely allow to find patterns between words and their associated model, and one should study how the model evolves when its associated word is given a specific transformation.

6 Conclusion

This project aimed at first using a neural network model to solve the spam detection problem, then analysing and interpreting that model using state-of-the-art XAI methods. After preliminary data pre-processing and Exploratory Data Analysis, a rather simple neural network based classifier was implemented, evaluated and compared with the Naive Bayesian algorithm to attest its robustness. In the second part of the project, two different XAI approaches were tested : linguistic correlation analysis and frozen weights method.

Linguistic correlation analysis allowed to visualize the impact of a specific neuron on the output of the model, whereas frozen weights aimed at describing the model in a vectorial space. Both allowed to display some relevant results and insights, but were still confronted to some limitations. For future works, the prolongation of this project should revolve around three axes:

- Improving the methods introduced here to overcome their limitations;
- Observing the changes in the outputs with different models or data sets;
- Comparing these methods with other on the subject.

References

[AbdulNabi and Yaseen2021] Isra’a AbdulNabi and Qussai Yaseen. 2021. Spam email detection using deep learning techniques. *Procedia Computer Science*, 184:853–858. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

[Almeida2011] Gomez Hidalgo J.M. Yamakami A. Almeida, T.A. 2011. Contributions to the study of sms spam filtering: New collection and results. *2011 ACM Symposium on Document Engineering (DOCENG’11)*, pages 1–9.

[Belinkov and Glass2019] Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72.

[Christopher D. Manning2008] Prabhakar Raghavan Hinrich Schütze Christopher D. Manning. 2008. *Introduction to Information Retrieval*.

[Harisinghaney et al.2014] Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora. 2014. Text and image based spam email classification using knn, naïve bayes and reverse dbscan algorithm. pages 153–155, 02.

[I. Androutsopoulos2003] E. Michelakis I. Androutsopoulos, G. Paliouras. 2003. Learning to filter unsolicited commercial e-mail.

[Mark Hopkins1999] George Forman Jaap Suermondt Mark Hopkins, Erik Reeber. 1999. Spambase data set.

[Shahariar et al.2019] G. M. Shahariar, Swapnil Biswas, Faiza Omar, Faisal Muhammad Shah, and Samiha Binte Hassan. 2019. Spam review detection using deep learning. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0027–0033.

[V. Metsis and Paliouras2006] I. Androutsopoulos V. Metsis and G. Paliouras. 2006. Spam filtering with naive bayes - which naive bayes? *3rd Conference on Email and Anti-Spam (CEAS 2006).*