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

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.

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.

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.

3.1 Spam detection

We will be testing different XAI methods with spam detection in mind as we will be training our models on few spam collection data sets.

3.1.1 History

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.

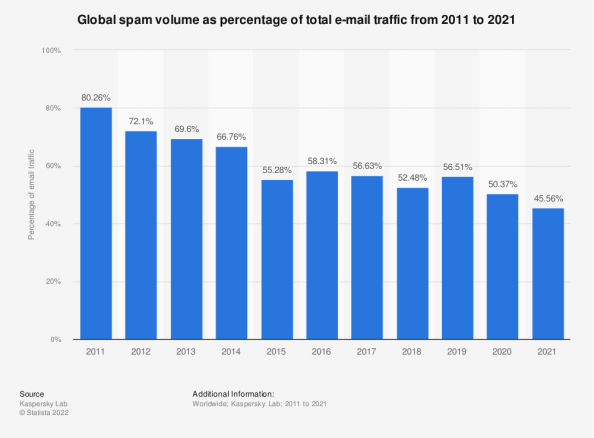


Figure 1: Evolution of the proportion of spam in the global e-mail traffic.

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

3.1.2 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).

3.2 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.

3.2.1 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.

(Dalvi, 2018)

4 Work Plan

4.1 Planification and timeline

The work plan that will be adopted is the following one.

1. Preliminary work (bibliographical research and work environment setup)
2. Implementation of a spam detection algorithm
3. Implementation of interpretation methods for neural networks
4. Research and Improvement
5. Report and presentation

4.2 Research goals

Here are the main research goals of the project:

- **Review** the recent methods for interpreting neural networks

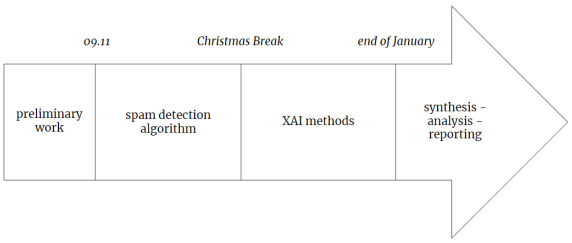


Figure 2: Timeline of the project realization.

- **Research** new development possibilities
- **Synthesize** and propose a new methodology for neural networks interpretation

4.3 Risk analysis

For this project, the main risks are the inability to complete the tasks in the given schedule and the irrelevance of the results obtained. These risks are at higher chance during the XAI phase since it is a very recent and broad field.

In the first case, there would be a need to restructure the work plan, lower the ambitions and focus on basic methods (only the first one for the XAI for instance).

In the second case, other methods would be researched, even if less complex or less detailed. For example, if the results of XAI methods can’t be exploited, more importance would be given to simple visualization methods.

5 Methodology

5.1 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

5.2 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.

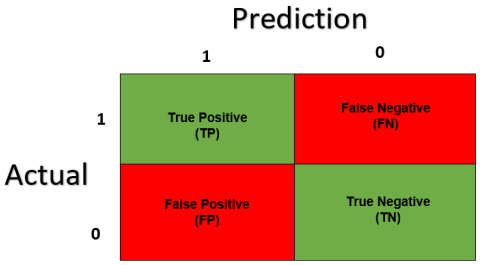


Figure 3: Computing the confusion matrix allows to access accuracy, recall and precision.

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.

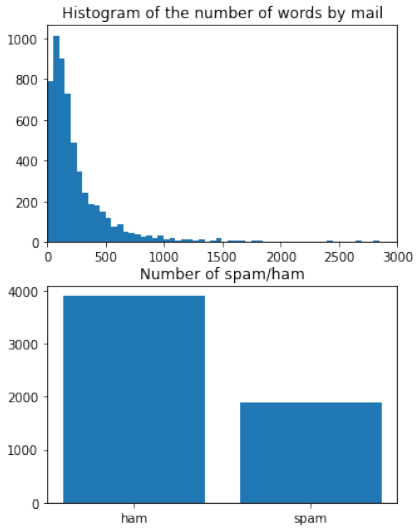
5.3 Data Sets

A preliminary research allowed to find several available data sets.

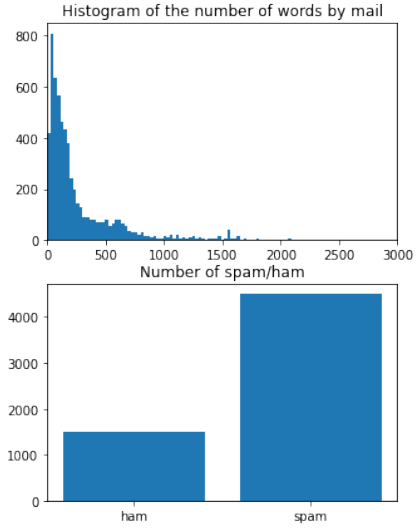
SpamAssasin - 2001 / 6047 instances
This data set has been created by the *Apache Software Fondation* to develop their spam filter software.

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

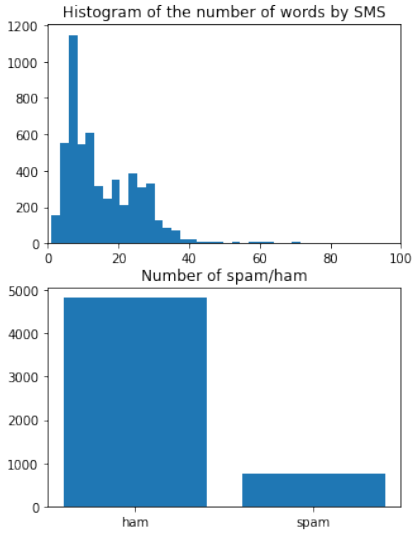
SMS Spam Collection - 2012 / 5574 instances
The data set was created and studied in the following article (Almeida, 2011). It contains SMS and not e-mails, but it will be used to compare the interpretations obtained on different formats of text.



(a) Spamassasin



(b) Enron



(c) SMS Spam Collection

Figure 4: Content of the data sets.

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.

5.4 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;

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.

[Dalvi2018] Durrani N. Sajjad H. Belinkov Y. Bau A. Glass J. Dalvi, F. 2018. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models.

[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)*.