Advanced Phishing Filter Using Autoencoder and Denoising Autoencoder

Samira Douzi
University Mohammed V
Faculty of Science
LRI. B.O. 1014, Rabat, Morocco
212634340703
samiradouzi8@gmail.com

Meryem Amar
University Mohammed V
Faculty of Science
LRI. B.O. 1014, Rabat, Morocco
212666315707
amar.meryem@gmail.com

Bouabid El Ouahidi
University Mohammed V
Faculty of Science
LRI. B.O. 1014, Rabat, Morocco
Bouabid.ouahidi@gmail.com

ABSTRACT
Phishing is referred as an attempt to obtain sensitive information, such as usernames, passwords, and credit card details (and, indirectly, money), for malicious reasons, by disguising as a trustworthy entity in an electronic communication [1]. Hackers and malicious users, often use Emails as phishing tools to obtain the personal data of legitimate users, by sending Emails with authentic identities, legitimate content, but also with malicious URL, which help them to steal consumer's data. The high dimensional data in phishing context contains large number of redundant features that significantly elevate the classification error. Additionally, the time required to perform classification increases with the number of features. So extracting complex Features from phishing Emails requires us to determine which Features are relevant and fundamental in phishing detection. The dominant approaches in phishing are based on machine learning techniques; these rely on manual feature engineering, which is time consuming. On the other hand, deep learning is a promising alternative to traditional methods. The main idea of deep learning techniques is to learn complex features extracted from data with minimum external contribution [2]. In this paper, we propose new phishing detection and prevention approach, based first on our previous spam filter [3] to classify textual content of Email .Secondly it's based on Autoencoder and on Denoising Autoencoder (DAE), to extract relevant and robust features set of URL (to which the website is actually directed), therefore the features space could be reduced considerably, and thus decreasing the phishing detection time.

CCS Concepts
Computing methodologies→Artificial intelligence • Computing methodologies→Information extraction • Computing methodologies→Machine learning • Computing methodologies→Supervised learning • Computing methodologies→Neural networks • Computing methodologies→Feature selection

Keywords
Autoencoder; Denoising Autoencoder; phishing; Spam-filter.

BDIOT2017, December 20–22, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5430-1/17/12…$15.00
https://doi.org/10.1145/3175684.3175690

1. INTRODUCTION
Email is the most popular way of communication of this era. It helps to transact information timely, rapidly and easily. For those reasons, Email has become a widely used by malicious users, which pretend to be legitimate senders. This attack method, commonly known as Email phishing, it can be considered as a subcategory of Spam, or even be jumbled with it [4] [5] .

The Anti-Phishing Working Group (APWG) Reports [6] that the total number of phishing attacks in 2016 was 1,220,523, a 65% increase over 2015. A Typical Phishing Email is sent at many potential victims' mailboxes, and usually comes with a clickable link, that victims receive effectively, convincing them to visit a fraudulent website, at which they are tricked into divulging sensitive information. Phishing attacks could have serious consequences for their victims, such as the loss of intellectual properties, sensitive customer information, financial loss and the compromise of national security [7]. [8]

Due that the Email is the most used Internet service nowadays, we proposed, in our previous work [3], a novel spam filter that try to overcome the weakness of the Bag of Word (BoW) approach, notably the fact that is ignoring the relationship between words. We achieved this by proposing a methodology that combines the representation of context of an Email, by using Paragraph Vector-Distributed Memory (PV-DM) approach, and the representation of its pertinent features selected by the well-known scheme called TF-IDF.

This methodology represents a more comprehensive filter for classifying Emails, due that it exploits the complementary that exists between words and their context. However, this filter can't detect an Email with legitimate content and malicious URL.

Furthermore, many studies have considered several Email Phishing features to detect phishing; they extracted various attributes from URL. The main limitation of those approaches is that, many of extracted features are evaluated without taking into account, whether they really are essential to identify phishing. [9] [10]. Therefore, it could lead to unnecessary computational cost in phishing detection, especially, with high flow of Emails.

In this paper, we proposed a new filter phishing constituted of our previous proposed spam filter, an Autoencoder and on DAE, that learn the structure of the data, and build by an unsupervised selection, a set of relevant features.

The proposed approach aims to facilitate data understanding, reduce feature space and decrease detection time.

This paper is organized as follows: Section 2 discusses the related works. While Section 3 introduces the basic theory of Autoencoder. Section 4 describes in detail our proposal approach for detecting Email phishing. And finally, in Section 5 we present conclusions and perspectives.

2. RELATED WORKS

Phishing is a growing problem on the internet today for both consumers and businesses. The most common approach used by attackers is creating a similar website, in order to capture personal information from consumers. A malicious website may look identical to an online bank or other trusted entity, in the interest of capturing passwords, social security numbers, account numbers, and other confidential information. A victim may not identify the malicious site until after the confidential information has been leaked.

Blacklisting is the most common anti-phishing technique used by modern web browsers, however, studies show that blacklist is not adequate enough to protect end users from new and emerging phishing web pages that appear in the thousands and quickly disappear every day. [11]

There are two major problems with blacklists: firstly, the IP address of the phisher and the malicious URL tend to change constantly to avoid the sender tracking or its identification. Secondly, blacklist fail to identify phishing URL in the early hours of a phishing attack, due that their update process is insufficiently fast [12].

On other hand, many studies have studied several Email features to detect phishing; it found that there are many factors that can distinguish the original legitimate URL from the forged URL; as Number of domains in the URL [13] [14], Using the IP address [15] [13] [16] [17], Using forms with ''Submit'' button [5] [18] [17], Number of dots [5] [18] [19], Hyperlink with image instead of visible text, and image URL based on IP address [5] [14] [17] etc. Maher Arborous et al [20] attempted a survey to identify the required Features which helps to improve the accuracy and the precision of detecting malicious URL. S. Shivaji et al prove that the number of features used in the detection engine has a direct impact in the processing time [9], and El-Khatib, K. shows that an exhaustive set of features may become the chokepoint of the Email system [10]. Fette and his colleagues used a technique that involves machine learning with 10 features, an anti-Spam tool and querying external sources (the whois service), to discover the "age of a domain" of the Email sender, or some URL in the Email body [5]. Such approach may increase considerably the time to evaluate each Email. The WHOIS service is only efficient whether the domain is recent, but in general the phisher hide himself under consolidated domains to avoid the detection by this type of query. Cook et al [21] propose using a classifier with 11 features, although the results reported a good detection rate, but some Features need clarification about its inclusion in the classification process. Zhang et al. [22] present CANTINA, content-based approach to detect phishing websites, based on the TF-IDF information retrieval algorithm and the Robust Hyperlinks algorithm. By using a weighted sum of 8 features (4 content related, 3 lexical, and 1 WHOIS-related), they show that CANTINA can correctly detect approximately 95% of phishing sites. However, downloading the actual web pages increases the potential risk of analyzing the malicious content on user's system.

Although some authors have used few features in detecting phishing emails, the main limitation of those approaches is that many features are evaluated without considering whether they really are pertinent to identify phishing. Therefore, it could elevate the classification error, or lead to unnecessary computational cost in phishing detection, especially with high flow of Emails. Moreover, there is no evaluation of correlations between features; such evaluation will allow obtaining the minimum set of distinct features with reliability similar to all features together, and assessing the importance of each one, so that we can affect to each feature a weight, which reflect its pertinence.

3. AUTOENCODER

The Autoencoder is an unsupervised learning method, which learns Features from the original input. The basic framework of Autoencoder [23, 24] is a neural network, which comprises an input layer, an output layer and at least one hidden layer. The aim of an Autoencoder is to transform inputs into outputs with the least possible amount of deviation. So it is usually used as an information compressor [25].Given the training samples $X=\{x_1, x_2 \ldots x_n\}$, such as for each sample $\mathbf{x_i}$, $x_i = \{ x_1,x_2,\ldots x_D\}^T$, weight Matrices $W_1 \in R^{d \times D}$, $W^T_1 \in R^{D \times d}$, and bias vectors $b_1 \in R^{d \times 1}$, $\hat{b}_1 \in R^{D \times 1}$,the encoder transforms the input vector $\mathbf{X}$ into a hidden representation $h=\{h_1,h_2,\ldots h_m\}$, such as for each $\mathbf{h_i}$, $h_i=\{h_1, h_2,\ldots..h_d\}$, and $(d<D)$, through a nonlinear activation function $\boldsymbol{f}$, such sigmoid or than function. :

$$h_i = f(W_1 \times x_i + b_1) \qquad (1)$$

The vector $\mathbf{h}$ is transformed back to a reconstruction vector $z=\{z_1,z_2, \ldots z_D\}$, such as for each $\mathbf{z_i}$, $z_i=\{ z_1,z_2,\ldots.z_D\}$, by the decoder as follow : $z_i = f(W_1^T \times h_i + \hat{b}_1)$ **(2) .**

The objective of Autoencoder is to enforce the output $\mathbf{z_i}$ to be as close as possible to the input $\mathbf{x_i}$, by minimizing the reconstruction error using Euclidean distance:

$$\min_{W_1, W_1^T, b_1, \hat{b}_1} \sum_{i}^{n} \left\| z_i - x_i \right\| \qquad (3)$$

4. OUR APPROACH

It should be noted that in this paper, we consider detection of the phishing Email as a classification problem. For any problem of this kind, we need a feature set and a classification algorithm. We have raw Emails as input and in training phase each Email is assigned a label as phishing or normal.

Our proposed phishing Email detection system works according to the architecture, depicted in Figure 1.

The architecture of the proposal phishing Email detection is comprised of six modules, which works as an assembly of tasks. The functions of each of the module are described as follow:

- **Input Emails**

This module is responsible for receiving Email contents as raw input. The Emails in this module contain each part of the Email, such as text and URL.

- **Content extractor**

This module of the architecture extracts the text with the URL of the Email. The reason for extracting only these two parts is that, they often describe the characteristics of the phishing Email.

- **Filter Based on PV-DM**

The Phishing Email mostly contains such a text that the receivers immediately turn to respond, by clicking on the links provided in the Email or send the crucial information in reply. However, research in the area usually focuses to classify phishing Emails by using the information available only on URL. [11]

Therefore, to analyze the text content of Email, we proposes as module in our phishing detecting architecture , the spam filter proposed in our previous article [3], that after analyzing the context of Email and its pertinent features , return a confidence value.

- **Features construction**

Once the set of URL extracted from Emails are available, feature construction engine builds up various feature-sets which are designed according to the experience and are found in various kinds of the phishing Emails.

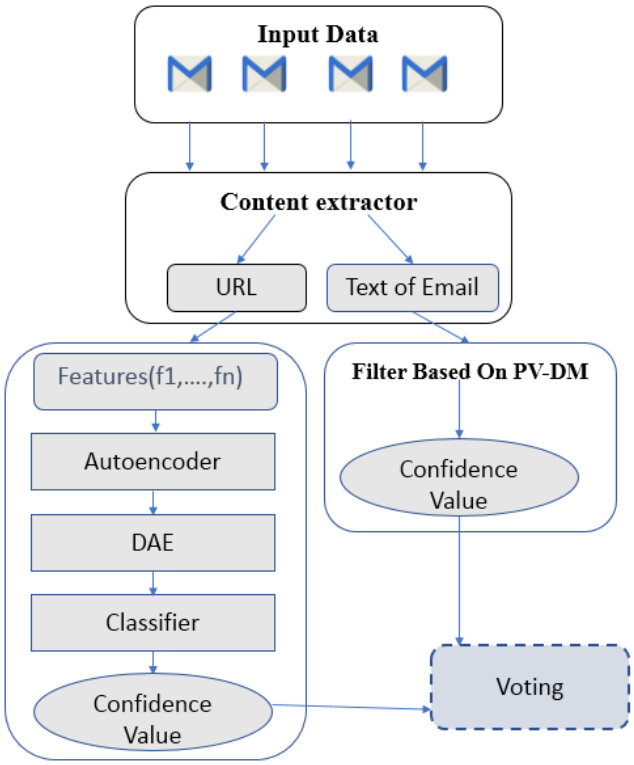

Figure 1: Proposed phishing Email detection system

- **Autoencoder**

The data of phishing URL contains a large number of features ranging up of tens of thousands of features [11] [9], which significantly elevate the classification error. Hence, to speed up classification algorithms, Autoencoder is employed to reduce the dimension of data, by an unsupervised algorithm to select the relevant features, which preserve the structure of original data.

- **Training of Autoencoder**

Given the training samples of Features extracted from URL, U = {U_1, U_2, …U_n }, The Autoencoder training consists of finding parameters W_1, W^T_1, b_1, $\hat{b}_1$, in order to enforce the output z_i to be as close as possible to the input x_i (see Figure 2), by minimizing the reconstruction error using Euclidean distance as the standard Autoencoder loss function. This corresponds to minimizing the following objective function:

$$\min_{W_1, W_1^T, b_1, \hat{b}_1} \sum_{i}^{n} \left\| z_i - x_i \right\| \quad (4)$$

one of the main characteristic of the Autoencoder is that it can be trained with unlabeled data, so once the hidden representation is learned, it can be taken as input to a supervised classifier that can be trained with a smaller labeled data, or is also possible to train other Autoencoder with these hidden representations.

- **Denoising Autoencoder**

Unfortunately, from time to time, the phishers create the new technique to circumvent filter, the new considered feature may even confuse the detection system, creating a point of uncertainty for the classifier. One of our goals is to create a robust filter to a partial change of the input. For that, we propose to introduce a Denoising Autoencoder (DAE) in our proposed system.

A Denoising Autoencoder (DAE) is a more recent variant of the basic Autoencoder consisting of only one hidden layer; it is trained to reconstruct a clean 'repaired' input from a corrupted version. [26]

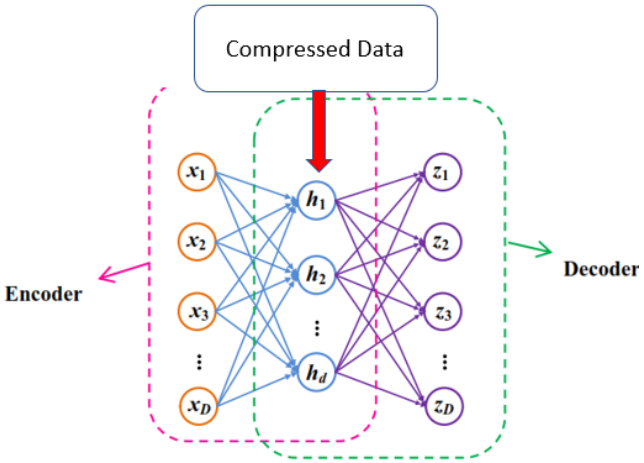

Figure 2: Training procedure of Autoencoder Training

- **Training of Denoising Autoencoder**

The procedure to train a Denoising Autoencoder is similar to Autoencoder training. The only difference here is that, the input is taken from the hidden representation of the first Autoencoder, and converted to a corrupted version by means of a corrupting function ϕ . So for each input **x**, a fixed number **C** of components are chosen at random, and their value is forced to 0, while the others are left untouched [27] , note that we can consider an alternative corrupting function. The Denoising Autoencoder will be trained to restore the initially input.

$$\hat{h} = \phi(h) \quad (5)$$

As with the basic Autoencoder the corrupted $\hat{h}$ input is mapped to a hidden layer:

$$\tilde{h} = g(W \times \hat{h} + b) \quad (6)$$

From which we reconstruct **y**

$$y = s(W' \times \tilde{h} + b') \quad (7)$$

Such as W, W' , are the weight Matrices, and b, b' the bias vectors of DAE.

The DAE is trained to reconstruct a repaired input from the corrupted version, by finding the parameters {W, W', b, and b'} which minimize the reconstruction error. This corresponds to minimizing the following objective function:

$$\sum_{i}^{n} \|y_i - h_i\| (8).$$
$$W,W',b,b'$$

In doing so; the learner must capture the structure of the input (Features), to reduce the effect of the corruption. In this way, more robust Features are learned compared to a first Autoencoder.

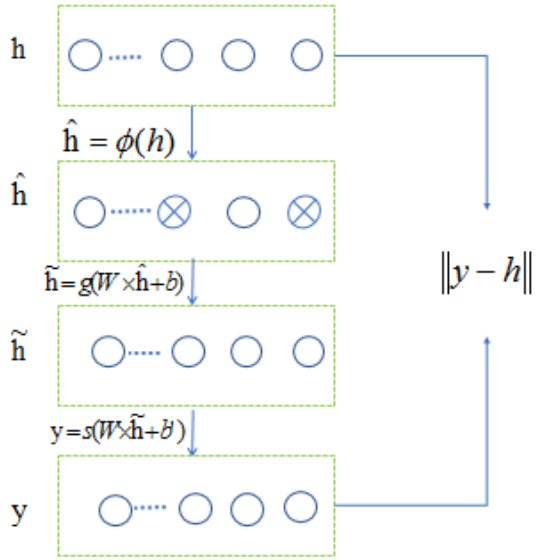

Figure 3: process of Denoising Autoencoder (DAE) training

- **Training of classifier**

The learned features by DAE will be the input for a classifier such as linear regression.

Given a test input $y_i, i = 1,.....,n$, and the class label c = {0, 1} (0 for phishing Email, 1 for Normal Email) the classifier logistic regression will estimate the probability:

$$p(c = j/y_i) \qquad \text{For each } j=0, 1$$

$$p(c = j/y_i, \theta) = \frac{e^{\theta_j^T y_i}}{\sum_{k=0}^{1} e^{\theta_k^T y_i}} (9)$$

Where $\theta = \{\theta_0, \theta_1\}$ are the parameters of the model. Given the training set $(y_i, c_i)_{i=1}^{n}, c_i \in (0,1)$ the solution of linear regression can be derived by minimizing the following formula:

$$\min_{\theta_0,\theta_1} \frac{-1}{n} \sum_{i=1}^{n} \sum_{j=0}^{1} I(c_i = j) \log \frac{e^{\theta_j^T y_i}}{\sum_{k=0}^{1} e^{\theta_k^T y_i}} (10)$$

Where I (c_i=j) is the indicator function, which equals to 1 only when the value of the statement is true, otherwise 0

After training the model, a new instance y_i is labeled to the class which has the biggest conditional probability:

$$c_i = \frac{e^{\theta_j^T y_i}}{\sum_{k=0}^{1} e^{\theta_k^T y_i}} \qquad (11)$$

- **Voting system**

Voting system is based on the confidence values returned by each filter (See Figure 1), the method consists of a simple voting, since confidence values are between zero and one, we calculate the average of both values, and the Email is labeled as Normal Email if the average is between *0.5* and *1* and phishing otherwise.

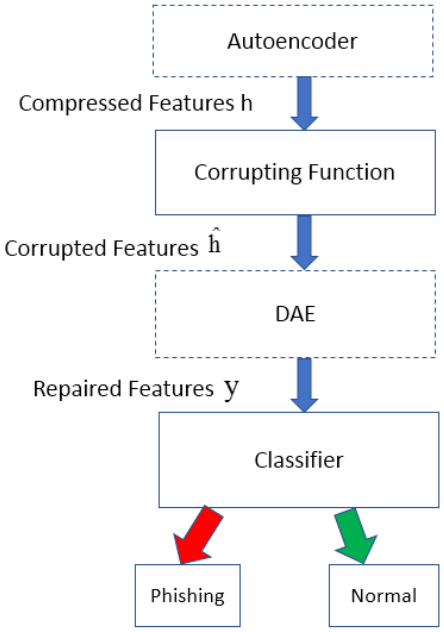

Figure 4: the global structure of URL filter based on Autoencoder and DAE

5. CONCLUSION AND PERSPECTIVE

In this paper, we have proposed a phishing detection architecture that increases the Email security, by a dual filter: one is for analyzing of textual content of Email and the second is for analyzing the suspect URL. To analyze the text Email, we propose using our previous spam filter [3], while for analyzing the URL, we have introduced a deep Autoencoder that finds a compressed representation of the Features.

In order to create a robust filter to a partial change of the input, the reduced set of Features is converted to a corrupted version; a Denoising Autoencoder is trained to reconstruct the repaired Features from this corrupted version.

Finally, we can feed these repaired Features directly to a classifier such as logistic Regression.

This methodology represents many advantages for classifying phishing Emails, due that: firstly, it exploits the information available on URL, and also on the text content of Email. Secondly it learns the structure of the data, and builds by an unsupervised Autoencoder, the relevant features of suspect URL. Third it could to reconstruct the Features even they are corrupted.

We are currently in the process of implementing this approach using python programming language, this will allow us to validate our work and produce pertinent experimental results.

6. REFERENCES

[1]. *Phishing attacks and countermeasures.* **Ramzan, Zulfikar (2010).** 2010, Handbook of Information and Communication Security. Springer. ISBN 9783642041174.

[2]. *Learning Deep Architectures for AI.* **Bengio, Yoshua.** s.l. : Foundations and Trends® in Machine Learning: Vol. 2: No. 1, pp 1-127. , 2009.

[3]. *Towards A new Spam Filter Based on PV-DM (Paragraph Vector-Distributed Memory Approach)***, Samira Douzi, Meryem Amar, Bouabid El ouahidi, Hicham Laanaya.** Science Direct ,Procedia Computer Science Volume 110, 2017, Pages 486–491 .

[4]. **E. El-Alfy, R. Abdel-Aal,.** Using GMDH-based networks for improved Spam detection and e-mail feature analysis. *Applied Soft Computing 11 (1) (2011) 477–488.*

[5]. **Ian Fette, Norman Sadeh,Anthony Tomasic.** Learning to Detect Phishing Emails. *International World Wide Web Conference, 2007, pp. 649–656.*

[6]. **http://www.apwg.org/resources/apwg-reports/.** *Phishing Activity Trends Report 4 th Quarter 2016.* s.l. : APWG.

[7]. *Phishing Attacks: Analyzing Trends in 2006.* **Ramzan, Z., & Wüest, C.** In Fourth conference on Email and Anti- Spam Mountain view: Citeseer, 2007.

[8] **Aaron, G.** The state of phishing . *Computer Fraud & Security* . 2010 (6) (2010) 5–8.

[9]. **S. Shivaji, E.J. Whitehead, R. Akella, K. Sunghun.** Reducing features to improve bug prediction . *IEEE/ACM International Conference on Automated Software Engineering (2009) 600–604.* 2009.

[10]. **El-Khatib, K.** Impact of feature reduction on the efficiency of wireless intrusion detection systems. *IEEE Transactions on Parallel and Distributed Systems 21 (8) (2010) 1143–1149.*

[11]. *LEARNING TO DETECT PHISHING URLs* . **al, Ram B. Basnet et.** s.l. : International Journal of Research in Engineering and Technology , Jun-2014, Vol. Volume: 03 Issue: 06 .

[12]. **Sheng, S.,Wardman, B.,Warner, G., Cranor, L., Hong, J. and Zhang, C.** An empirical analysis of phishing blacklists,. *In Proceedings of the CEAS'09, 2009.*

[13]. *Detection of phishing attacks: a machine learning approach , Soft Computing Applications in Industry (2008) 373–383.* **R. Basnet, S. Mukkamala, A. Sung,.**

[14]. *Obtaining the threat model for e-mail phishing. Appl. Soft Comput. J. (2011),.* **C.K. Olivo, et al.**

[15]. *Online detection and prevention of phishing attacks* . **J. Chen, C. Guo,.** s.l. : Communications and Networking in China (2006) 19–21.

[16]. *Profiling phishing e-mails based on hyperlink information , International Conference on Advances in Social Networks Analysis and Mining (2010) 120–127.* **J. Yearwood, M. Mammadov, A. Banerjee,.**

[17]. *Analysis of Phishing Attacks and Countermeasures* . **Biju Issac, Raymond Chiong and Seibu Mary Jacob.** s.l. : at www.arxiv.org., 2006.

[18]. *Detecting Malicious URLs in E-mail- An Implementation ,2013 AASRI Conference on Intelligent systems and control, Procedia 4 (2013) 125-131.* **Dhanalakshmi Ranganayakulu, Chellappan C.**

[19]. *Efficient prediction of phishing websites using supervised learning algorithms* . **Santhana Lakshmi V, Vijaya MS.** s.l. : International Conference on Communication Technology and System Design 2011,Procedia Engineering 30 (2012) 798 – 805.

[20]. **Maher Aburrous, Hossain, M.A., KeshavDahal and FadiThabtah.** "Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies. *Cognitive Computing,Vol. 2, pp. 242-253* . 2010.

[21]. **D. Cook, V. Gurbani, M. Daniluk, Phishwish: a stateless phishing filter using.** Phishwish: a stateless phishing filter using minimal rules , Lecture Notes in Computer Science (2008) 182–186.

[22]. *CANTINA: a content-based approach to detecting phishing web sites.* **Y. Zhang, J. Hong, L. Cranor.** s.l. : In Proc. 16th Int. Conf. World Wide Web, WWW"07 Banff, Alberta, Canada, 2007, pp. 639-648. .

[23]. *Representation Learning via Semi-supervised Autoencoder for Multi-task Learning.* **al., Fuzhen Zhuang at.** s.l. : EEE International Conference on Data Mining, 2015.

[24]. *Unsupervised Feature Extraction with Autoencoder Trees.* **Ozan úIrsoy, Ethem Alpaydõn.** s.l. : Neurocomputing (2017), doi:10.1016/j.neucom.2017.02.075.

[25]. *A novel deep autoencoder feature learning method for rotating machinery fault diagnosis.* **Shao Haidong, Jiang Hongkai,Zhao Huiwei, Wang Fuan.** s.l. : Mechanical Systems and Signal Processing 95 (2017) 187–204.

[26]. *Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition.* **Jun Deng, Student Member, IEEE, Zixing Zhang, Florian Eyben, Member, IEEE, and Björn Schuller, Member, IEEE.** s.l. : IEEE SIGNAL PROCESSING LETTERS, VOL. 21, NO. 9, SEPTEMBER 2014.

[27]. *Extracting and Composing Robust Features with Denoising Autoencoders.* **al, Pascal Vincent et.** s.l. : Proceedings of the 25 International Conference ence on Machine Learning, Helsinki, Finland, 2008.