Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 110 (2017) 486–491

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Symposium on Emerging Inter-networks, Communication and Mobility (EICM 2017)

Towards A new Spam Filter Based on PV-DM (Paragraph Vector-Distributed Memory Approach)

Samira Douzi*, Meryem Amar,Bouabid El Ouahidi, Hicham Laanaya

Mohammed V University, Faculty of Sciences, L.R.I.
B.O. 1014, Rabat, Morocco

Abstract

The increasing volume of emails has led to the emergence of problems caused by unsolicited email, commonly referred to as Spam. One of the most commonly presentation used in Spam Filter is the BoW (Bag-of-words). However, this approach has a number of weaknesses, mainly the fact that the word order is lost; hence different emails can have the same representation since the same words are used, and it ignores the relationship between words, which can lead to poor performance. This paper proposes a new Spam filter based on PV-DM (Paragraph Vector-Distributed Memory) in order to overcome the limitations of the BoW representation.

*Keywords:*spam filter ; Paragraph Vector-Distributed Memory; word embedding.

1. Motivation

Email is one of the most popular, fastest and efficient ways to exchange information. However, the wide use of email has led to the emergence of problems caused by unsolicited email messages, commonly referred to as Spam. According to annual reports, the average of spam sent per day increased from 2.4 billion in 2002 to 300 billion in 2010[1], and according Symantec Intelligence Report, the global ratio of spam in email traffic is 71.9%[2]

* Corresponding author. Tel.: 212634340703
Email address: samira.douzi@yahoo.fr

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.06.130

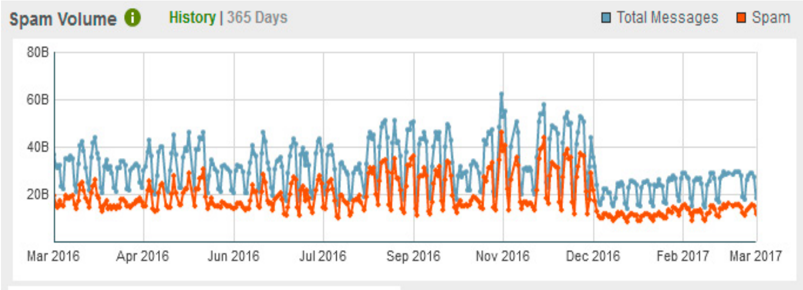

Fig. 1. Global spam and email volume[3]

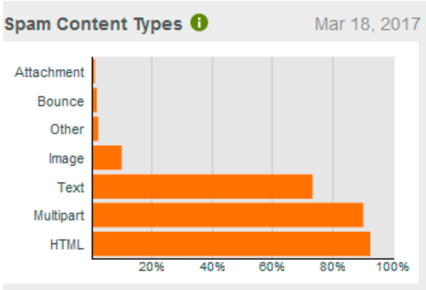

Fig .2. spam content types[3]

The high volume of undesirable Spam messages is beginning to diminish the reliability of email. This causes information security problems[4], such as the stealing of sensitive information, phishing, etc. The economic impacts of Spam have led some countries to adopt legislation to combat it, although this is limited by the fact that such messages can be sent from outside their jurisdiction[5, 6].

In order to tackle this problem, Spam filters are used. These filters analyse the direct textual content of the email message, as well as additional information. One of the most popular approaches used in Spam filters is the bag-of-words BoW representation, also known as the vector-space model[5].

Given a set of terms $\mathbf{F= \{t_1, t_2 \ldots\ldots, t_n\}}$, BoW represents an email text **E** as a n-dimensional feature vector $\mathbf{X=\{x_1 \ldots\ldots, x_n\}}$ where the value of $\mathbf{x_i}$ is given as a function of the occurrence of $\mathbf{t_i}$ in **E**, depending on the representation of the features adopted. The features are generally given as single words occurring in messages used for training.

However, this representation introduces some problems: firstly, the word order is lost; hence different emails can have the same representation since the same words are used. Secondly, it is assumed that the initially selected features will always be representative for the classification of a message. This has been proven to be false because Spammers are continually trying to create new ways to overcome Spam filtering. For instance, spammers obscure content and disguise certain terms that are very common in Spam messages, e.g., by writing "*fr33*" instead of "*free*", or "*mon3y*" instead of "*money*". This is an attempt to prevent the correct identification of these terms by Spam filters[7].

Motivated by above, this paper reports a novel spam filtering approach, to facilitate accurate and efficient spam classification. It can exploit the embedded information enclosed in the context of emails, as well as its relevant features, this method consists of representing each email as a continuous distributed vector through the use of PV-DM model. This technique is inspired by the recent work in learning vector representation[8, 9, 10, 11].

This paper is organized in the following way. In section 2, we briefly review some of the previous work that attempts to address the limitations of the BoW. In section3, we present the basic concept of PV-DM model. In section 4, we present in detail our proposed methodology. A conclusions and directions for future works end this paper in section 5.

2. Previous works

The limitations of feature representation BoW (see Fig 3) method are: Firstly, bag-of-words model encode every word in the vocabulary as One-Hot-Vector, meaning that for a vocabulary of size $|V|$, each word is represented by a $|V|$ dimensional sparse vector with *1* at the index corresponding to the word and *0* at every other index. As vocabulary may potentially run into millions, the BoW model leads to the high dimensional indexing problem. Secondly, it ignores the conceptual similarity between terms, as example, the words 'car' and 'automobile' are often used in the same context. However, the vectors corresponding to these words are orthogonal in BoW model, which can lead to poor performance. Thirdly, while modelling sentences using Bag-of-words, the order of words in the phrase is not respected. Ex: "This is Free" and "Is this Free" have exactly the same vector representation.

Some efforts have been made to address these limitations. For example, Elisabeth Crawford et al[12]showed that using phrase-based representation can be used to increase performances of email classifiers. Matthew Chang and Chung Keung Poon[13] studied the use of phrases as the basic features in the email classification problem. They employ three different classifiers, namely, a naive Bayes classifier and two k-NN classifiers using Term Frequency-Inverse