Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 833–843

Procedia Computer Science

www.elsevier.com/locate/procedia

21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

# A Topic-Based Hidden Markov Model for Real-Time Spam Tweets Filtering

Mahdi Washha[a,*], Aziz Qaroush[b], Manel Mezghani[a], Florence Sedes[a]

[a] *IRIT Laboratory, University of Toulouse, Toulouse, France*
[b] *Department of Electrical and Computer Engineering, Birzeit University, Ramallah, Palestine*

**Abstract**

Online social networks (OSNs) have become an important source of information for a tremendous range of applications and researches such as search engines, and summarization systems. However, the high usability and accessibility of OSNs have exposed many information quality (IQ) problems which consequently decrease the performance of the OSNs dependent applications. Social spammers are a particular kind of ill-intentioned users who degrade the quality of OSNs information through misusing all possible services provided by OSNs. Social spammers spread many intensive posts/tweets to lure legitimate users to malicious or commercial sites containing malware downloads, phishing, and drug sales. Given the fact that Twitter is not immune towards the social spam problem, different researchers have designed various detection methods which inspect individual tweets or accounts for the existence of spam contents. However, although of the high detection rates of the account-based spam detection methods, these methods are not suitable for filtering tweets in the real-time detection because of the need for information from Twitter's servers. At tweet spam detection level, many light features have been proposed for real-time filtering; however, the existing classification models separately classify a tweet without considering the state of previous handled tweets associated with a topic. Also, these models periodically require retraining using a ground-truth data to make them up-to-date. Hence, in this paper, we formalize a Hidden Markov Model (HMM) as a time-dependent model for real-time topical spam tweets filtering. More precisely, our method only leverages the available and accessible meta-data in the tweet object to detect spam tweets exiting in a stream of tweets related to a topic (e.g., #Trump), with considering the state of previously handled tweets associated to the same topic. Compared to the classical time-independent classification methods such as Random Forest, the experimental evaluation demonstrates the efficiency of increasing the quality of topics in terms of precision, recall, and F-measure performance metrics.


Peer-review under responsibility of KES International

*Keywords:* Hidden Markov Model, Social Spam, Real-Time, Twitter.

## 1. Introduction

Online Social Networks (OSNs) have an enormous popularity over the Internet because of the wide range of services that they provide for their users. For example, the most popular OSNs such as Twitter, and Facebook have exceeded billions of registered users and millions of daily active users[1]. The key point of the OSNs is their dependency on users as primary contributors in generating and posting information. Users' contributions might be exploited in different positive ways such as understanding users' needs, and analyzing users' opinions for election

* Corresponding author
*E-mail addresses:* mahdi.washha@irit.fr (Mahdi Washha)., aqaroush@birzeit.edu (Aziz Qaroush)., Mezghani.Manel@gmail.com (Manel Mezghani)., florence.sedes@irit.fr (Florence Sedes).

1877-0509 
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.075

purposes. However, the simplicity and flexibility of using OSNs in addition to the absence of an effective restrictions on content posting action have exposed different information quality problems such as social spam, and information overload[2]. Indeed, these characteristics have subjected OSNs to different attacks by ill-intentioned users, so-called social spammers, to post spam content. Social spammers intensively post nonsensical content in different contexts (e.g, topics) and in an automated way[3]. For example, posting a tweet talking about "how to gain 100$ in 5 minutes" under the "#BBC" topic is a spam tweet since such a tweet has no relation to the given topic. Thus, social spammers have a wide range of goals when publish a spam content in OSNs, summarized in[4]: (i) spreading advertisements to generate sales; (ii) disseminating porn materials; (iii) publishing viruses and malware; (iv) and creating phishing websites.

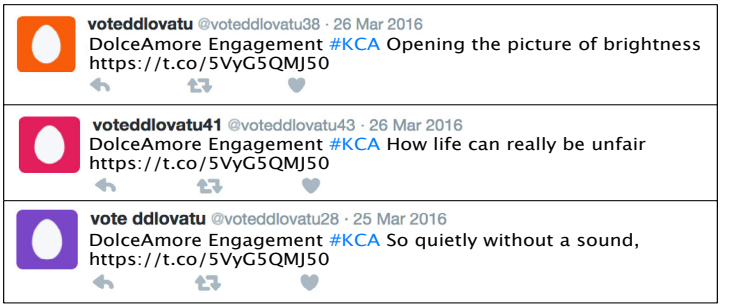


Figure 1. An example of three correlated spam tweets posted in a consecutive way by three different spam accounts.

**Motivation and Problem.** Since OSNs have many information quality problems, in this work, a particular issue related to the social spam problem in Twitter platform has been addressed. More precisely, we address the problem of filtering out the spam tweets that might exist in a stream of tweets related to a Twitter topic to increase the topics content quality, with taking into consideration the real-time aspect of the filtration process. The proposed solution, has been integrated with our team researches on social networks. The research interests of our team addressing many issues related to OSNs such as tweet summarization, event detection[5], social profiling[6], profiles enrichment[7], and socio-semantic communities detection[8], where Twitter platform has been adopted as a source of information in most of them. Thus, experimenting and working on a high quality of Twitter data (Tweet content) is an indispensable step to achieve high performance results in our team researches. Besides the information quality requirement, some research topics, such as tweets summarization and events detection, require real-time spam tweets filtering.

As Twitter is not immune towards the social spam problem[1], a set of methods has been introduced in the literature for detecting spam campaigns and individual spam accounts[4,3,9,10,11,12,13,14], with little effort spent for individual spam tweets detection[4,15,16]. These efforts mainly exploit supervised machine learning methods combined with the features extraction concept to produce binary classifiers using annotated data-sets. However, some of these methods such as campaign and account based methods are "not" suitable for real-time filtration because their features require an additional information from Twitter's servers. The only legal way to retrieve these additional information is by using REST APIs[1] which are provided by Twitter for developers and researchers. However, Twitter imposes limitations and constraints (e.g., limiting the number of calls to a time slot) on using REST APIs, decreasing the applicability of such methods in the a real-time way. Moreover, exploiting graph-based features such as node betweenness, and sender-receiver distance require an exponential number of REST APIs calls to extract their values.

Most of the features that are used at tweet-level detection such as number of words are light in computation and thus they are suitable for real-time spam tweets detection. However, given the fact that social spammers are dynamic in their content and strategies[3], these light features are not strongly discriminant among spam and non-spam tweets. Also, the combination of these weak features is not necessary to produce robust binary classification models, since social spammers are easily manipulate in these features value. The straightforward and trivial solution to address such a problem is designing new light features having enough discriminant power among spam and non-spam tweets. However, this solution is not possible since the tweet content is limited to 140 characters and the simple available meta-data about its user (e.g., username attribute) increases the difficulty to design new robust light features. Beyond the feature design level, the approaches followed in building spam classification models are time-independent learning algorithms (e.g., Random-Forest, Support Vector Machine) in which the learning and classification steps are performed without considering the state of previous classified instances. Also, updating and tunning the classification models (i.e., Model parameters) that use those learning methods require a wide range of training and validation to

[1] https://dev.twitter.com/rest/public