# Chapter 4
# Thwarting Spam on Facebook: Identifying Spam Posts Using Machine Learning Techniques

**Arti Jain**
*Jaypee Institute of Information Technology, India*

**Reetika Gairola**
*Jaypee Institute of Information Technology, India*

**Shikha Jain**
*Jaypee Institute of Information Technology, India*

**Anuja Arora**
*Jaypee Institute of Information Technology, India*

## ABSTRACT

*Spam on the online social networks (OSNs) is evolving as a prominent problem for the users of these networks. Spammers often use certain techniques to deceive the OSN users for their own benefit. Facebook, one of the leading OSNs, is experiencing such crucial problems at an alarming rate. This chapter presents a methodology to segregate spam from legitimate posts using machine learning techniques: naïve Bayes (NB), support vector machine (SVM), and random forest (RF). The textual, image, and video features are used together, which wasn't considered by the earlier researchers. Then, 1.5 million posts and comments are extracted from archival and real-time Facebook data, which is then pre-processed using RStudio. A total of 30 features are identified, out of which 10 are the best informative for identification of spam vs. ham posts. The entire dataset is shuffled and divided into three ratios, out of which 80:20 ratio of training and testing dataset provides the best result. Also, RF classifier outperforms NB and SVM by achieving overall F-measure 89.4% on the combined feature set.*

## INTRODUCTION

In the today's world the change in the Internet technology has led us to the usage of different Online Social Networks (OSNs)[1] (Andreassen et al., 2016; Brown et al., 2008; Egele et al., 2017; Panicker & Devadas, 2015), also known as the changing web. The out bursting popularity of these OSNs have at-

DOI: 10.4018/978-1-5225-5097-6.ch004

tracted huge number of users to use their platform which results into the sharing and storing of users' personal information on these networking sites. These networks help the users to interact, exchange and collaborate with their social-circle. These OSNs are also helping its users to communicate with their social community and keep their users updated with different domains such as news, active learning, job searching and web application development etc. Such vital information has aroused the interest of spammers[2] to take the advantage of the trust among users to deceive them for spammers (Adewole et al., 2017) own benefits. Facebook OSN is currently among one of the leading OSN present across the world and having over 2.05 billion[3] monthly active users. Facebook is around five times greater than its next greatest partner Twitter. A survey report[4] shows that out of 5,173 adults suggested that 30% of people get their news from Facebook, while only 8% receive news from Twitter and 4% from Google Plus. The users of Facebook not only uses Facebook for communicating with their friends but also for keeping regular updates about what is happening around the globe.

Spammers at present are discovering different ways to reach out to the users of the OSNs for spreading spam messages (Bhat & Abulaish, 2013; Prieto et al., 2013) and thereby, making the OSNs as vulnerable and exposed targets. Mostly, these spam messages are sent in high volume so that they can influence large amount of users in a short span of time. Moreover, these messages reduce the memory of the inboxes. These messages are targeted to specific audience or can be used to perform tricks such as phishing, identity theft (Gao et al., 2012). Apparently, spam which was earlier in the form of text containing irrelevant information for the users is now currently it is being noticed as it being spread using images and videos too. To do so, spammer evades the programmed filters using different techniques such as obfuscating keywords, wrapping long urls, and using image or video instead of textual content.

In this chapter, a methodology to segregate spam vs. ham post[5] from Facebook OSN is provided by combining the textual, image and video features using three supervised Machine Learning (Shalev-Shwartz & Ben-David, 2014) techniques, namely- Naïve Bayes (NB) (Lee et al., 2010), Support Vector Machine (SVM) (Shalev-Shwartz & Ben-David, 2014) and Random Forest (RF) (Breiman, 2001) using RStudio®[6]. Our methodology comprises of various stages. Firstly, data extraction is done from the Facebook posts which contain text, image and video posts followed by the data pre-processing (DP) stage. DP stage consists of stop words removal, stemming, lemmatization, photo pre-processing and url link blacklisting. The next stage consists of relevant features extraction for the identification of whether a post is spam or not. In this stage, a total of 30 features mare extracted from Facebook which includes *"status type", "created time", "updated time", "name of the photo"* etc. Then the data is shuffled and split into different training and test data ratios i.e. 60:40, 70:30 and 80:20 respectively. Then ML techniques are applied to train the three classifiers (NB, SVM and RF). Further, using RStudio these classifiers are used in the testing phase to predict whether a given post is spam or ham. Finally, the classifiers are evaluated on the basis of standard metrics- precision, recall and F-measure.

## BACKGROUND

Spam and ham posts classification of Facebook users' posts is a noticeable and prominent research issue which has been addressed by various researchers in their work. It is observed that these researchers have spam campaigns using individual features- textual features (Meligy et al., 2015), image features (Gao et al., 2012) and video features (Benevenuto et al., 2009). On one end, researchers primarily have worked on the textual posts. On the other end, few researchers have tried to classify based on image or video