Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits

Carl Sabottke Octavian Suciu Tudor Dumitraș
University of Maryland

A Appendix

A.1 Features used for classification

The four categories of features we use are: Twitter Words, Twitter Traffic, CVSS Information and Database Information.

Table 2 lists the word features from the Twitter Text category, along with their mutual information with respect to both the real-world and the public proof-of-concept ground truth data sets. The "X" marks in the table indicate that the respective words did not end up in the final feature sets after the thresholding step.

Table 1 lists the features in the other three categories, together with the type of each individual feature.

No.	Feature Name
Twitter Traffic	
1	Number of tweets
2	Number of users with minimum 460 followers
3	Number of users with minimum 400 friends
4	Number of retweets
5	Number of status replies
6	Number of tweets added to favorites by someone
7	Average number of hashtags per tweet
8	Average number of URLs per tweet
9	Average number of user mentions per tweet
10	Number of verified accounts among posters
11	Average age of accounts among posters
12	Average number of tweets of accounts among posters
CVSS Information	
1	CVSS Score
2	Access Complexity
3	Access Vector
4	Authentication
5	Availability Impact
6	Confidentiality Impact
7	Integrity Impact
Database Information	
1	Number of references from NVD
2	Number of distinct sources in the references from NVD
3	'BUGTRAQ' found among NVD references
4	'SECUNIA' found among NVD references
5	'allow' found in NVD summay
6	NVD last modified date - NVD published date
7	NVD last modified date - OSVDB disclosed date
8	Number of tokens in OSVDB title
9	Current date - NVD last modified date
10	'OSVDB' found among NVD references
11	'code' found in NVD summay
12	Number of OSVDB entries for CVE
13	Vulnerability Category assigned in OSVDB
14	Vulnerability Category assigned by our regular expressions
15	First vendor from NVD
16	Number of distinct vendors from NVD
17	Number of affected products from NVD

Table 1: Feature summary

Keyword	MI Wild	MI PoC	Keyword	MI Wild	MI PoC
advisory	0.0007	0.0005	ok	0.0015	0.0002
beware	0.0007	0.0005	mcafee	0.0005	0.0002
sample	0.0007	0.0005	windows	0.0012	0.0011
exploit	0.0026	0.0016	w	0.0004	0.0002
go	0.0007	0.0005	microsoft	0.0007	0.0005
xp	0.0007	0.0005	info	0.0007	X
ie	0.0015	0.0005	rce	0.0007	X
poc	0.0004	0.0006	patch	0.0007	X
web	0.0015	0.0005	piyolog	0.0007	X
java	0.0007	0.0005	tested	0.0007	X
working	0.0007	0.0005	and	X	0.0005
fix	0.0012	0.0002	rt	X	0.0005
bug	0.0007	0.0005	eset	X	0.0005
blog	0.0007	0.0005	for	X	0.0005
pc	0.0007	0.0005	redhat	X	0.0002
reading	0.0007	0.0005	kali	X	0.0005
iis	0.0007	0.0005	0day	X	0.0009
ssl	0.0005	0.0003	vs	X	0.0005
post	0.0007	0.0005	linux	X	0.0009
day	0.0015	0.0005	new	X	0.0002
bash	0.0015	0.0009			

Table 2: Mutual information provided by the reduced set of keywords with respect to both sources of ground truth data.