A Framework for Real-Time Spam Detection in Twitter

Himank Gupta, Mohd Saalim Jamal, Sreekanth Madisetty and Maunendra Sankar Desarkar
Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India
Email: [cs16mtech01001, cs16mtech11024, cs15resch11006, maunendra]@iith.ac.in

1. Introduction

Due to the increasing popularity of Twitter, this platform influences more number of spammers to generate spam tweets. Twitter platform experienced a **355% growth of social spam** during last 4 years.

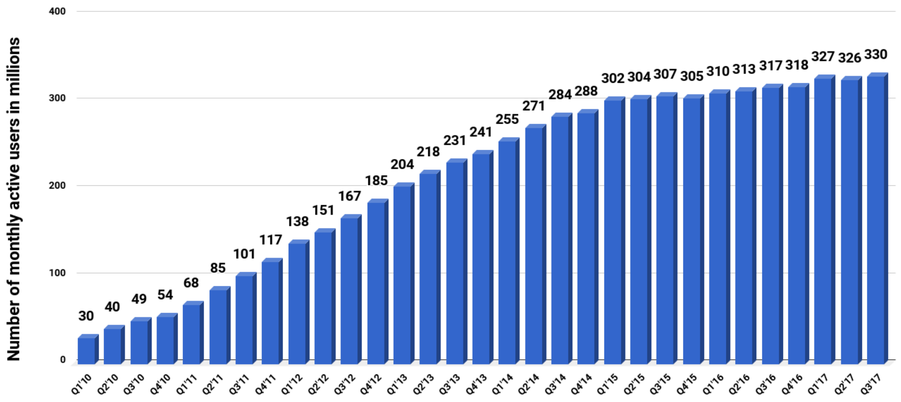


Figure 1: Growth of users on Twitter platform in last 7 years [a]

[a] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users

2. Motivation

1. It is crucial to detect Twitter spam as soon as possible in real-time because 90% of users might visit a new spam link before it gets blocked by the blacklist [1].
2. We need to choose lightweight features that should be feasible to process a large number of tweets in very less time.

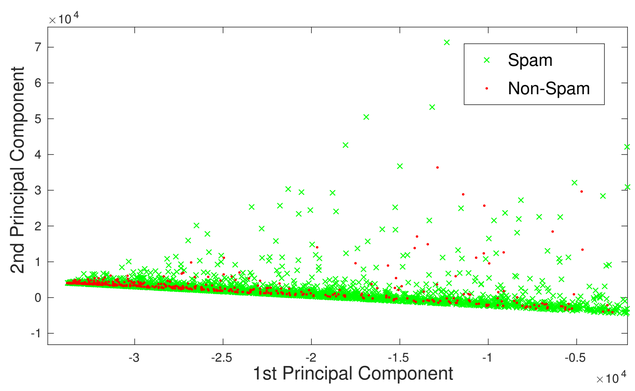


Figure 2: Scatter plot of dataset [2] showing distribution of two classes namely, spam(x) and non-spam(.)

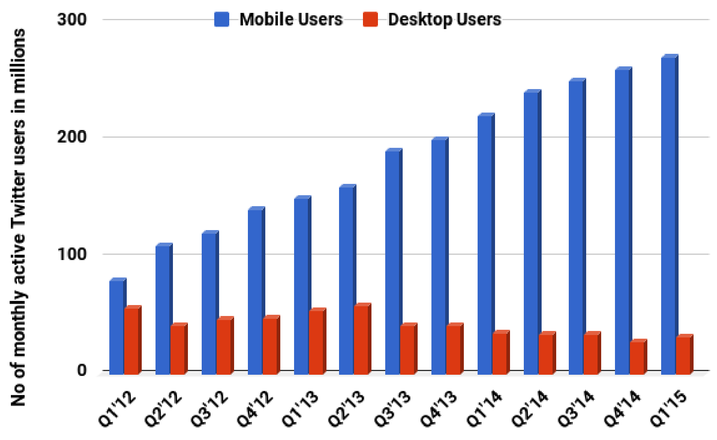


Figure 3: Use of Twitter platform on different devices [a]

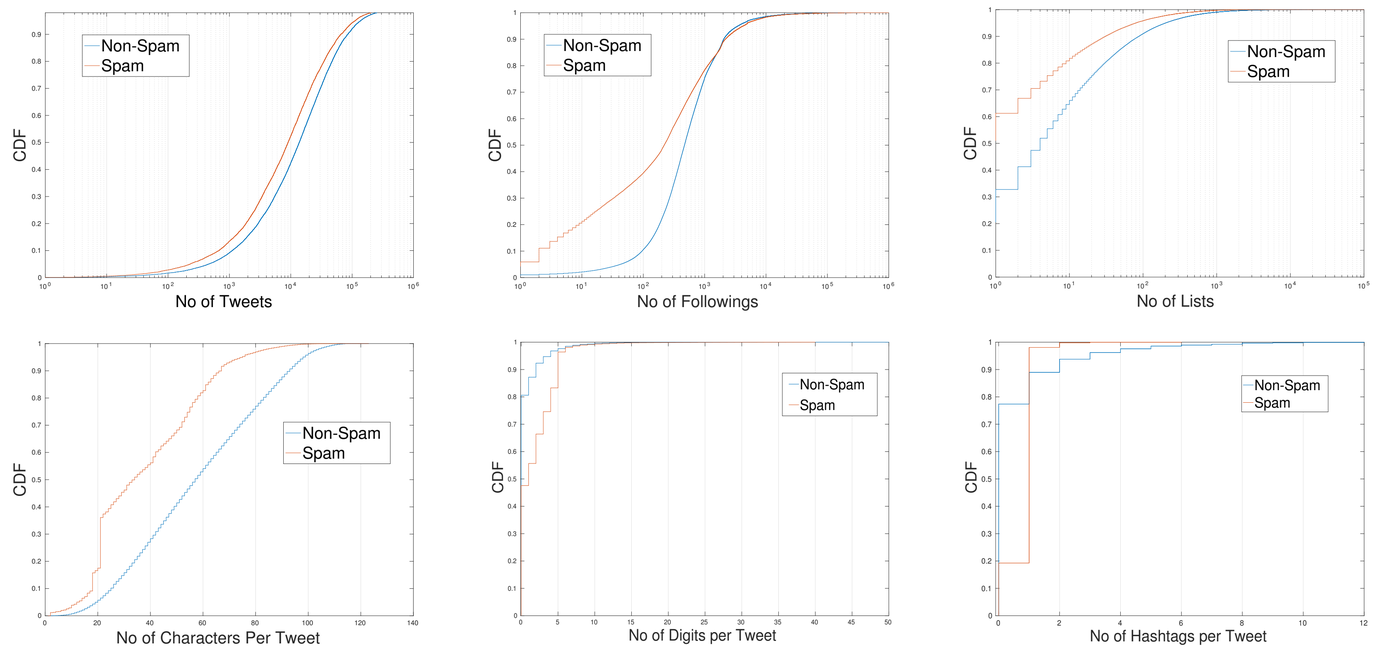
[a] https://www.statista.com/chart/1520/number-of-monthly-active-twitter-users/

3. Extracted Features

Feature Name	Description
account age	The age (days) of an accounl since its creation until the time of sending the most recent tweet
no_followers	The number of followers of this Twitter user
no_followee	The number of followee/friends of this Twitter user
no_userfavourites	The number of favourites this Twitter user received
no_lists	The number of lists this Twitter user added
no_tweets	The number of tweets this Twitter user sent
no_retweets	The number of retweets this tweet
no_hashtags	The number of hashtags included in this tweet
no_usermentions	The number of user mentions included in this tweet
no_urls	The number of URLs included in this tweet
no_chars	The number of characters in this tweet
no_digits	The number of digits in this tweet
no_non-ASCII_characters	The number of non-ASCII characters in this tweet

4. Cumulative Distribution Functions of Features

We investigated each feature's characteristics of differentiating spam and non-spam tweet. Following figures shows the Cumulative Distribution Function (CDF) of three user based features and three message based features. Analysis of these features has showed us their discriminative power to detect Twitter spam.



5. Proposed Work

$\mathcal{U}_{\mathcal{S}}$ = Collection of unique words in the spam tweets' text. $\mathcal{U}_{\mathcal{NS}}$ = Collection of unique words in the non-spam tweets' text.
For each word $\mathcal{T}$ in $\mathcal{U}_{\mathcal{S}}$ and $\mathcal{U}_{\mathcal{NS}}$ we calculate the following probablity values:

$$\mathsf{P}(\mathcal{T}|\mathcal{U}_{\mathcal{S}}) = \frac{\#\text{ of Spam tweets that contain } \mathcal{T}}{\text{total }\#\text{ of Spam tweets}} \qquad \mathsf{P}(\mathcal{T}|\mathcal{U}_{\mathcal{NS}}) = \frac{\#\text{ of Non-Spam tweets that contain } \mathcal{T}}{\text{total }\#\text{ of Non-Spam tweets}}$$

We calculate the information gain $\gamma_{\mathcal{T}}$ for each word $\mathcal{T}$ as follows:

$$\gamma_{\mathcal{T}} = \left| \frac{\mathsf{P}(\mathcal{T}|\mathcal{U}_{\mathcal{S}})}{\mathsf{P}(\mathcal{T}|\mathcal{U}_{\mathcal{NS}})} \times \log_{10} \frac{\mathsf{P}(\mathcal{T}|\mathcal{U}_{\mathcal{S}})}{\mathsf{P}(\mathcal{T}|\mathcal{U}_{\mathcal{NS}})} \right|$$

We use top-30 words based on information gain $\gamma_{\mathcal{T}}$ along with the lightweight features described in section 3.

Table 1: Top 10 Words

Spam Words	Non-Spam Words
harvested	rain
tribez	asleep
coins	rather
collected	college
unfollower	fell
openfollow	follback
inspi	dinos
build	bullshit
smurf	child
brainy	couch

Table 2: Classification of Example Tweets

S.No	Sample Tweet with Feature Set	With Bag-of-Words Model	Without Bag-of-Words Model
1.	I've collected 12,293 gold coins! http://t.co/MXyllUOlZa #android, #androidgames, #gameinsight (1944,11,19,0,0,13134,0,3,0,1,21,5,1)	Spam	Spam
2.	Also, please go out and vote for your local councillors today #publicserviceannouncement (3419,923,753,674,33,29769,0,1,0,0,50,0,0)	Non-Spam	Non-Spam
3.	Get Unlimited Followers!!!! https://t.co/EKqMEvvmGF (5,0,9,0,0,1,0,0,0,1,21,0,4)	Spam	Non-Spam

6. Results

Table 3: Sampled Features Set

Feature-Set	Sampling Method	Ratio (Spam:Non-Spam)
1	Use 43 features to train a model	1:2
2	Use Bag-of-Word to select features in libsvm format	1:2
3	Use Chao Chen's [2] feature-set for comparison	1:1

Table 4: Performance evaluation on different Feature set

	Feature-set- 1		Feature-set- 2		Feature-set- 3	
Classifier	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
SVM with Kernel	86.18	85.95	84.28	83.88	79.9	79.1
Neural Network	90.56	91.65	-	-	71.25	72.15
Gradient Boosting	75.81	85.84	-	-	81.26	82.69
Random Forest	75.39	86.25	-	-	93.6	92.9

7. Conclusions & Future Work

1. We present a novel framework for real-time spam detection in Twitter using top-30 words along with the lightweight features that outperformed the existing work [2] by 18%.
2. We will keep on updating our Bag-of-Words model based on new spam tweets to mitigate the ***"Spam-Drift"*** problem.
3. We are developing **smart phone application** and **browser extension** to detect Twitter Spam in real-time as 80% of Twitter users access Twitter via their mobile devices.
4. We observe in our dataset that 79% of spam tweets contain a malicious link. So we will also perform the ***URL crawl mechanism*** along with ***Frequent Pattern Mining*** of tweets' text to distinguish Twitter spam in real-time.

8. References

[1] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 243–258, New York, NY, USA, 2011. ACM.

[2] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou. 6 million spam tweets: A large ground truth for timely twitter spam detection. In *2015 IEEE International Conference on Communications (ICC)*, pages 7065–7070, June 2015.