Topic-based engagement analysis of an online social service

Nripesh Trivedi,
Mathematical Sciences, IIT (BHU) Varanasi

Abstract- The content on the web is increasing at an explosive rate. This makes it important to generate interpretations of web content in a way that could facilitate building of relevant systems. Keeping this notion in focus, an analysis is carried out to understand and generate interpretations of engagement of users with respect to topics discussed on an online social service [5] (here a hacker forum). Since this could be the first research of its kind, results are presented in a way that may be easily interpreted and used for carrying out further research. A number of existing machine learning techniques and models are used in this research. There is a section that points out the usefulness of these results in encouraging user participation.

Keywords: User, Engagement, Topic modeling, Online social service, Hacker Forum

1 INTRODUCTION

Can we use analysis of the context popular on an online social service (here hacker forum) to leverage its user base? Is it possible to obtain the overall direction of a hacker forum (an online social service [5]) by just looking at the title of the threads? Even if one is successful in coming up with reasonable answers to previous two questions, is it possible to use this information to obtain significant insights for the companies and industries concerned with this online social service? These are the questions which motivate this research. Previous works in the context of hacker forums range from threat detection to identification of key sellers [8]. A previous work on hacker forums has also used machine learning techniques to find out topics discussed over IRC channel in [1]. Despite the fact that number of attempts have been made to research hacker forums, no previous work has attempted to conduct a research on hacker forums by using engagement analysis coupled with Topic modeling. In this research, data obtained from Wilders Security Forum (https://www.wilderssecurity.com/) is used to first apply Topic modeling and then carry out an engagement analysis over it. A Topic modeling algorithm is applied to the title of the threads to find out about the topics discussed on the forum. Since the titles of threads are concise they are capable of giving a summary of the discussion in as few words as possible. This analysis on a hacker forum would help in bringing out the interests of community of hackers and this information could be further used to develop more relevant ethical hacker forums and websites. Moreover, this research could also help businesses and organizations that have cyber-security concerns in keeping themselves updated with the overall direction of interests of the hacker's community.

2 DATA-SET DESCRIPTION

A data-set consisting of over 27,000 rows (each row corresponding to a thread in the forum) is used. The original data-set consists of the following attributes

Table 1: Original Data-set

Name of the Attribute
Title of a thread
Total number of replies to a thread
Total number of views for a thread

Title of the threads in the original data-set are used in Topic modeling.

Table 2: Data-set used in Topic modeling (data-set 1)

Name of the Attribute
Title of a thread

3. TOPIC MODELING (LATENT DIRICHLET ALLOCATION)

In this section, the usage of Latent Dirichlet Allocation (LDA) to extract topics from the dataset is described. In order to evaluate the models and subsequently choose the best model, perplexity is used[2]. LDA is a common technique to find out about the topics discussed and has been successfully used in a variety of fields. One such application of LDA has been made in [7]. The language used for implementation is R.

3.1 Pre-processing of Data

For the purpose of stop-word removal, POS(Part-of-Speech) tagging is used. Apart from the nouns and the adjectives, all other parts of speech are filtered out. This is a common technique to remove stop words [7]. This helps in reducing the document size as well as in removing the words that could make the application of LDA to result in finding out irrelevant topics. A number of libraries provide the functionality of POS tagging. For the analysis presented here, POS tagger from Stanford CoreNLP suite is used. In addition to this, a number of functions provided by ‘tm’ package in R are used as well. These functions include removal of numbers, punctuations, stemming and removal of specific words that may be commonly found in hacker forums.

3.2 Application of LDA

For the purpose of this research, LDA algorithm provided in the ‘topicmodels’ package (in R) is used. To find out the best model in this case, the dataset is partitioned into two parts, the training set and the test set. 60% of the dataset is used for training while the rest 40% is used for testing. Trained models are evaluated on test set by using perplexity. The lower this measure, the better fit the trained model is for the test set [2].

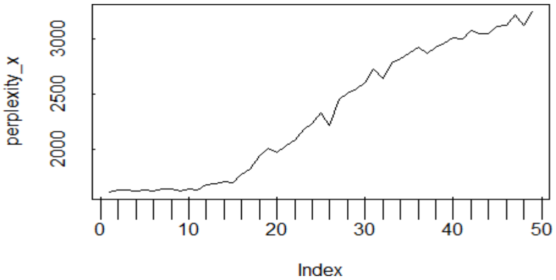


Figure I: Plot of perplexity against the number of topics

Using perplexity, the model suited to the data-set was found. In other words, the measure gave the optimal value of ‘k’ i.e. the number of topics. Another very important measure that needs to be considered in the application of LDA is hyper parameter, alpha. The commonly used value of this parameter is 50/k [7]. If one needs to restrict the assignment of individual documents (here thread titles) to fewer topics, than a lower value of alpha should be used [7]. Using this notion, a lower value of alpha (5/k) is used rather than the commonly used (50/k) [7]. From figure 1, it can be observed that perplexity increases rapidly after k=15, so a suitable value of 'k' should be less than 15. The optimality of the topics found is also determined by the interpretability of the topics. In other words, the words corresponding to the topics found should convey some meaning. Keeping this notion in focus, k=10 is chosen. The topics found are given in the table 3.

Table 3: List of topics with Top 5 words in each topic

Topic 1	help, spyware, error, browser, returnnil
Topic 2	new, malware, use, program, usb
Topic 3	question, image, drive, file, software
Topic 4	backup, free, install, boclean, home
Topic 5	attack, good, desktop, time, rootkit
Topic 6	window, partition, update, microsoft, rollback
Topic 7	linux, virus, virtual, release, version
Topic 8	sandboxing, remove, cant, rid, tool
Topic 9	log, trojan, system plea review
Topic 10	Problem, ubuntu, computer, truecrypt, worm

4. Engagement Analysis

Titles of threads are categorized into topics using Latent Dirichlet Allocation. Total numbers of threads belonging to each topic are counted. This counting gives the total number of threads in each topic Total number of replies and views for a given topic are also counted. The resulting data-set used is shown below:

Table 4: Data-set 2

Name of the Attribute
Total number of threads for a topic
Total number of replies to a topic
Total number of views for a topic

The 10 topics described in table 3 could be categorized broadly into 3 categories. The three categories are described in the table 5. The topics in each category are described in table 6 following table 5.

Table 5: List of categories with description of each category

Category 1	Seeking help as infected by malware, virus
Category 2	Discussing and promoting malware, spyware
Category 3	Seeking help for software, installation

Table 6: List of categories with topics in each category

Category 1	topic1 topic 9 topic 10
Category 2	topic 2 topic 7 topic 5
Category 3	topic 3 topic 4, topic 6 and topic 8

In [5], authors propose three engagement dimensions; initiation, interaction and loyalty to measure user engagement [5]. The social elements [5] of an online social service [5] (here hacker forum) can be grouped along three dimensions of user engagement [5]. Total numbers of threads belonging to each category are counted. This counting gives the total number of threads in each category Total number of replies and views for each category are also counted. The social elements used in engagement analysis are:

- Total number of threads of a category
- Total number of replies for a category
- Total number of views for a category

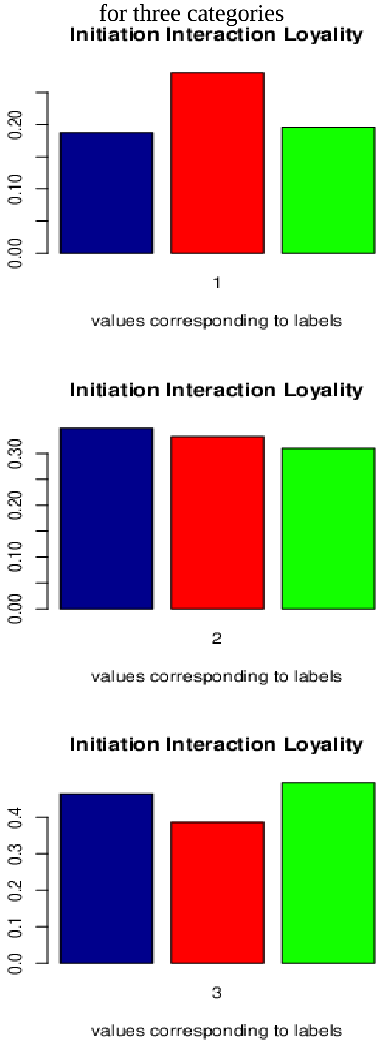
Following the description of each dimension in [5[, the above three social elements [5] could be assigned an engagement dimension as shown in table 7 below.

Table 7: Dimensions of user engagement with social elements in each dimension

Initiation	Total number of threads for a category
Interaction	Total number of replies to a category
Loyalty	Total number of views for a category

Using min-max normalization, each social element is re-scaled to a number between 0 and 1. For example, total number of threads for each category are assigned a numeral between 0 and 1. Since only one social element is present in each category, the name of the social element can be replaced with the name of the corresponding engagement dimension for convenient representation of user engagement. Since the dimensions are used for measuring user engagement and in case of single social element, this replacement does not affects the representation. In discussion below, the relevance of this representation is explained. The results obtained after applying min-max normalization to each social element in each category are shown in figure 2 below.

Figure 2: Representing engagement using engagement dimensions for three categories



Topics in category 1 experience more engagement along the dimension of interaction in comparison to the dimensions of initiation and loyalty. This implies that users are more likely to interact if the topics belongs to category 1. Topics in category 2 experience a decreasing trend in engagement on moving along the engagement dimensions, namely, initiation, interaction and loyalty. It implies that users tend to start threads relevant to topics in this category with a lot more enthusiasm that slowly fades. Topics in category 3 experience less engagement along the dimension of interaction in comparison to initiation and loyalty. This implies that users are less likely to interact if the topics belongs to category 3. The behavioral patterns of users described above is general to any online social service since the social elements in Table 7 are also general to any online social service.

5. Conclusion

In this research, broad categories of topics were found that users are interested in discussing over hacker forums. The application of this research is to increase user participation. User participation may be enhanced by engaging them with topics that they are most likely to interact with as shown in figure 2.

References

[1] Victor Benjamin, Weifeng Li, Thomas Holt, and Hsinchun Chen. 2015. Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In Intelligence Security Informatics (ISI), 2015 IEEE International Conference on. IEEE, 85–90.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022.

[3] Carina Jacobi, Wouter van Atteveldt, and Kasper Welbers. 2016. Quantitative analysis of large amounts of journalistic texts using Topic modeling. Digital Journalism 4, 1 (2016), 89–106.

[4] Weifeng Li, Hsinchun Chen, and Jay F Nunamaker Jr. 2016. Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System. Journal of Management Information Systems 33, 4 (2016), 1059–1086.

[5] Trivedi, N., Asamoah, D. A., & Doran, D. (2018). Keep the conversations going: engagement-based customer segmentation on online social service platforms. *Information Systems Frontiers*, *20*(2), 239-257.