Predicting political motivated distributed denial of service attacks from Twitter

Web Science and Engineering, IN4252

Dominik Lucas Harz
4513649
d.l.harz@student.tudelft.nl

ABSTRACT

Distributed Denial of Service (DDoS) attacks represent a majority of attacks occurring to websites and services. About 23% of the Internet's traffic is created by DDoS attacks [9]. They are comparably easy to execute and are partly used as a political tool for censorship and protest [4][1]. As shown in [12] social media can be used to track DDoS attacks. Furthermore, in [1] Twitter is used to create an early warning system for large disturbances of the Internet (i.e. natural disasters, political DDoS attacks). Within this paper a prediction of political motivated DDoS attacks based on information gathered from Twitter is introduced. Within the experiments tweets between December 25, 2015 to January 7, 2016 are monitored to determine DDoS attacks. It is shown that X DDoS attacks occurred during that time period. This is verified with data from a public DDoS monitoring service and internet routing updates. 50% of the predictions can also been seen from these other sources.

1. INTRODUCTION

Denial of service attacks are targeted to block the normal operation or communication of a specific systems to others. Distributed denial of service attacks (DDoS) are executed from multiple systems to target a specific or multiple systems at once [13]. In 2015 each DDoS attack caused around $115,000 of revenue loss for the affected organization [9]. Moreover, the number of DDoS attacks is increasing since the last 10 years [8]. There are several approaches undertaken to minimize the impact of DDoS attacks [4]. Furthermore, research is focussed on predicting DDoS attacks before they actually occur. These include social models that are used to predict contagions based Twitter and include DDoS attacks[1]. Also, extracting general security relevant events from Twitter is analysed [12].

Within this paper an approach to determine DDoS attacks from Twitter is presented, that are caused or related to a political motivation. DDoS can be seen as a form of political protest in a digital world [1]. Examples include the Low Orbit Ion Canon that was used to block websites of organizations opposing Wikileaks[1] and the use of the High Orbit Ion Cannon against the US Department of Justice after the shut-down of the website Megaupload[2].

In section 2 System Design a general approach how to predict DDoS attacks from Twitter is described. This is followed by experiments on two data sets from Twitter in section 3 Experiments. Results are presented and discussed in section 4. The article is concluded in section 5.

2. SYSTEM DESIGN

The system is designed to predict DDoS attacks from tweets based on information which is contained within the tweet, the user tweeting and the time the tweet is send. Therefore, the systems consists of four modules and two databases as presented in figure 1. In the first module, the Twitter streamer, the official Twitter Streaming API [3] is used to stream tweets and store the tweets in a database. In the second module, the Statistics and Sentiments module, the tweets are analysed to reveal potential insights. As described in [14] certain features of a tweet are used as indicators of its relevance to a certain topic. The following features are covered in this paper:

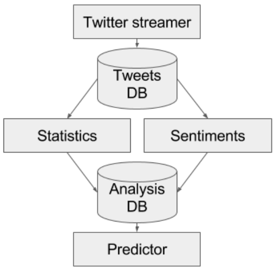

Figure 1: Overview of system design with streamer, statistics, sentiments and predictor module

length: This counts the number of characters used within one tweet. As presented in [14] the length of a tweet does not give particluar relevance to a certain topic. Here it is used to determine if it can add value to the given classification problem of political motivated DDoS attacks.

hasURL: This registers if and how many URLs are contained in a single tweet. If tweets contain a URL they are

[1] http://www.bbc.com/news/technology-11968605

[2] http://money.cnn.com/2012/01/19/technology/megaupload

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

generally more relevant to a certain topic[14].
hasHashtag: Hereby it is registered if a tweet contains one or multiple hashtags. Although similar to the length feature it might not be discriptive to general problems, experiments with this features are conducted..
hasEntity(ddos): To determine the entities related to DDoS a similar approach as in [7] is used. Wikipedia articles including "Denial-of-Service attack" [4] and further linked documents are analysed for word occurrences. Common words are filtered and the remaining words or combination of words represent the entities determining DDoS type entities. These words include for example: attack, ddos, denial-of-service, victim, flood, network, dos, system, packet, attacker, agent and tool.
hasEntity(politics): Similarly, the entities for politics are extracted from Wikipedia articels relating to politics, parties and current political situations in the UK. These entities include for example: party, government, parliament, united, ireland, scottish, labour, uk, eu, brexit, northern and kingdom.
sentiment: Furthermore, a simple implementation of sentiment analysis is realized with dictionaries of positive and negative words [6]. In each tweet the number of positive and negative words is counted and divided by the total number of words expressing an opinion. In case the number of positive words is greater then the number of negative words the sentiment score of a tweet is between 0 and 1. In the opposite case the sentiment score will be between -1 and 0. Neutral tweets receive a sentiment score of 0. This implementation is used to keep the analysis of a tweet relatively fast and have an unsupervised classification. How this can be enhanced is discussed in section 4.
The predictor module is implemented to determine which tweets indicate a potential political motivated DDoS attack. As it is not practical to label the individual tweets, the classification is based on an unsupervised classifier. In multiple cases tweets are classified with Support Vector Machines (SVM) to determine either the sentiment [3], topics [10] or user attributes[11]. In this case the SVM classifies the tweets in regular and outliers, that potentially indicate politically motivated DDoS attacks. To determine the outliers a one-class SVM with non-linear kernel is trained on tweets and tested with a second set of tweets. Outliers are represented by tweets containing entities relating to DDoS, politics or a combination thereof.

3. EXPERIMENTS

The system is implemented on a regular commodity hardware with an Intel i5 processor and 8 GB of RAM on a Linux Mint 17 operating system. The databases are realized with PostgreSQL 9.5.1[5]. The modules are written in Python. The twitter streamer uses the tweepy library[6], the sentiments module utilizes nltk[7] and the predictor utilizes the scikit-learn library[8]. Plots are created with matplotlib[9]. From December 25, 2015 to January 7, 2016 tweets from the UK[10] are streamed and saved to a database. Hereby a total number of 4,546,682 are received. Their timely distribution is presented in figure 2. Figure 2 shows that during the day most tweets are sent with an increasing number of tweets towards the evening. During night time around 01:00 to 05:00 only very little tweets are sent. There are spikes on the Christmas days (December 25, 2015), New Year's Eve (December 31,2015 to January 1, 2016) and on Epiphany (January 6, 2016). These tweets are then analysed to determine the features as described in section2. The following table 1 represents the number of features observed in the collected tweets. With these features the SVM is trained to classify normal tweets and to detect outliers. As this is done with a one class SVM a threshold has to be defined to determine outliers. In the experiments thresholds between 1% and 15% are evaluated in 1% steps. In each case the classifier is trained on 50% of the data and tested on the other 50%. It can be noted that the larger the number of samples the closer the error gets to the defined threshold. As the data is multidimensional, there is no plot available. The identified dates by the classification are December 26, 2015, December 30, 2015, January 1, 2016 and January 7, 2016. The results will be further. However, it is possible to extract the dates and number of tweets, which are assigned as outliers in relation to particular features. This is presented in figure 3. These results are discussed in the upcoming section.

Table 1: Number of features present in streamed tweets

Feature	Number
hasURL	1,221,950
hasHashtag	772,846
hasEntity(DDoS)	44,136
hasEntity(politics)	136,409
sentiment (+)	2,166,708
sentiment (-)	2,275,530

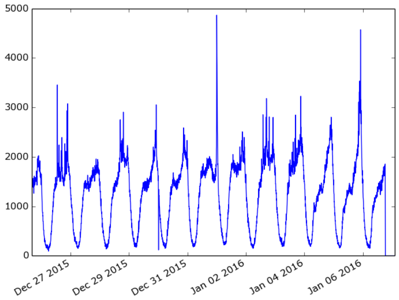

Figure 2: Number of tweets within the UK from December 25, 2015 to January 7, 2016 in intervals of 5 minutes (n = 4,546,682)

4. RESULTS AND DISCUSSION

To determine the validity of these results the error of classification can not be used as the data is not labelled. Instead two approaches are used. Firstly, this data is verified from

[4] https://en.wikipedia.org/wiki/Denial-of-service
[5] http://www.postgresql.org/
[6] http://www.tweepy.org/
[7] http://www.nltk.org/
[8] http://scikit-learn.org/stable/
[9] http://matplotlib.org/
[10] [-6.37988,49.871159 to 1.76896,55.811741]

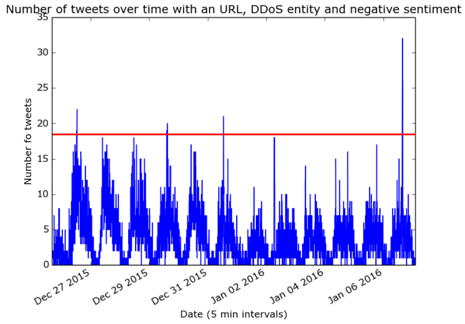

Figure 3: Number of tweets in relation to hasURL, hasDDoS and negative sentiment including the 5% outliers in intervals of 5 minutes.

the Digital Attack Map service. During the recorded time around one major DDoS attack within the UK occurred that is publicly monitored by Digital Attack Map. Digital Attack Map is a joint project by Google Ideas and Arbor Networks to monitor and report on major DDoS attacks[11]. This attack occured on the morning of December 30, 2015 on BBC and is one of the largest DDoS attacks that ever occurred[12]. Hence, the classified DDoS attack on this date seems to be correct.

Furthermore, RIPE offers raw data of routing updates that take place on the internet backbone in London. If a DDoS attack occurs normally routes are update to reroute traffic as to hinder congestion of internet lines [12]. Figure 4 presents the routing updates. Also here the DDoS attack on December 30, 2015 can be seen. Furthermore on January 7, 2016 routing changes are executed which can be an indicator, that on this date a DDoS attack has taken place. However, by crawling through news and media sites this could not be determined for attacks in the UK. It is though possible that an attack has been executed from the UK to another country, which can explain the missing news relating DDoS attacks on this date for the UK.

The other two dates can not be confirmed by other sources. The amount of tweets on January 1, 2016 in the morning is generally very large and thus might result in the spike in the data. For this the algorithm needs to be adjusted to cope with exceptional dates, where the amount of tweets is very high and thus the likelihood of having tweets containing entity related words or phrases is also very high.

For the four identified cases no evidences of a political motive was found from other sources than the tweets. Thus, it is possible that those DDoS attacks are performed out of other reasons. Therefore, it needs to researched if the here implemented solution can be generalized to serve as a general prediction model for DDoS attacks. Also, the entities identified as relevant for political motivation certainly need to be more detailed and further specified. As there is no political motive present, there should be no DDoS attacks displayed. However, the other features outweigh the hasEntities(politics) and thus the DDoS attacks are recognized.

The approach of classifying tweets can be further enhanced. Data can be labelled to enable the use of supervised or semi-supervised classifiers. Thereby, other common classifiers such as k-Nearest Neighbour, Parzen, Naive Bayes, other SVMs or different types of neural networks can be evaluated to classify single tweets. Thereby, a near-time solution can be created which can generate alarms after reaching a threshold of a certain number of tweets that indicate a DDoS attack.

Additionally, splitting the data into two subsets for training and testing is suboptimal as it might lead to over-fitting to the data. This can be optimized by using cross-validation to split the original dataset and run the classification multiple times to determine an average error.

A main weakness of this approach is that a recognition of particular entities is not possible. In order to determine the potential victim or attacker the tweets can be further analysed for named entities [2]. This needs to be included in the modules and can be an enhancement to the overall solution. Moreover, the current implementation of the sentiment analysis can be optimized. This can be achived by calculating sentiment analysis with the help of [5] or [3]. Currently, emoticons are not taken into account, which loses valuable part of the information. Also, correlation between different words is not taken into account, which can lead to a false classification of the sentiment.

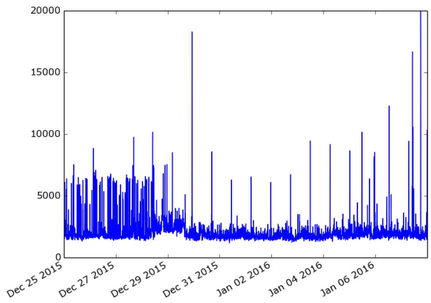

Figure 4: RIPE routing updates for the Internet backbone in London from December 25, 2015 to January 7, 2016.

5. CONCLUSION

This paper shows that DDoS attacks can be predicted from Twitter. In its current implementation it is however not very sensitive to the political motive or intention. Nevertheless, 50% of the identified DDoS attacks are verified from the Digital Attack Map or the RIPE data. The other identified attacks can be false positive. This is very likely for the New Year's Eve case. However, for the remaining occurrence this can also be a DDoS attack that is not reported or covered in the media.

It needs to be further researched if this approach can be enhanced to cover other motivations for DDoS attacks. Also, it needs to be researched if the model can be generalized to generic DDoS detection based on Twitter. For this a longer period of time should be monitored. To cope with the amount of data an initial filtering is required or a large scale system. The analysis of 4.5 million tweets requires a considerable amount of time on regular hardware.

[11] http://www.digitalattackmap.com/about/

[12] http://www.telegraph.co.uk/news/bbc/12075679/BBC-website-crashes-and-Twitter-goes-into-meltdown.html

6. REFERENCES

[1] R. Colbaugh and K. Glass. Leveraging sociological models for prediction ii: Early warning for complex contagions. In D. Zeng, L. Zhou, B. Cukic, G. A. Wang, and C. C. Yang, editors, *ISI*, pages 72–77. IEEE, 2012.

[2] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva. Analysis of named entity recognition and linking for tweets. *CoRR*, abs/1410.7182, 2014.

[3] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6, 2009.

[4] F. Lau, S. Rubin, M. Smith, and L. Trajkovic. Distributed denial of service attacks. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 3, pages 2275–2280 vol.3, 2000.

[5] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: Named entity recognition in targeted twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[6] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, New York, NY, USA, 2005. ACM.

[7] C. Lu, W. Lam, and Y. Zhang. Twitter user modeling and tweets recommendation based on wikipedia concept graph, 2012.

[8] J. Mirkovic and P. Reiher. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev.*, 34(2):39–53, Apr. 2004.

[9] PonemonInstitute. The cost of denial-of-services attacks. 2015.

[10] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: Supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 247–250, New York, NY, USA, 2012. ACM.

[11] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA, 2010. ACM.

[12] A. Ritter, E. Wright, W. Casey, and T. Mitchell. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 896–905, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[13] S. M. Specht and R. B. Lee. Distributed denial of service: Taxonomies of attacks, tools, and countermeasures. In *ISCA PDCS*, pages 543–550, 2004.

[14] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant for a topic? volume 838 of *CEUR Workshop Proceedings*, pages 49–56. CEUR-WS.org, 2012.