Detecting Spammers on Twitter Using Machine Learning Techniques: A Review

Connor Goddard

Department of Computer Science, Aberystwyth University
Aberystwyth, Ceredigion, SY23 3DB
Email: clg11@aber.ac.uk

I. Introduction

In recent years, the world has witnessed the explosion of social media and with it a dramatic change in the way people communicate with each other in their daily lives. Much of this change can be attributed to the enormous progress that has been accomplished in technologies related to the Internet and mobile computing, and this in turn has led to a significant re-shaping of communication behaviours observed across the current 21st century generation [1]. Today the popularity of social media continues to rise steadily, with many businesses and governments also beginning to recognise the immense reach that this technology can provide with respect to their target audiences.

Amongst all of the most popular social networking sites, Twitter has rapidly emerged as one of the best performers in terms of user growth, reporting an increase of 5% in user registrations between 2013 and 2014 alone [2]. In comparison to many other social networking platforms (e.g. Facebook or MySpace), Twitter enforces a highly simplified interaction mechanism that includes: limiting the length of any single item of content (known typically as a '*tweet*') to a total of 140 characters; and allowing for the possibility of non-mutual relationships between account holders [3].

With the rise in recognition for Twitter as a source of real-time news distribution and discussion, spammers have begun to capitalise on large news events as a means of rapidly boosting their audience size. Capable of operating in real-time, this new era of unsolicited content distribution brings about entirely new levels of challenge and complexity in the fight against unwanted advertising, and the circulation of indecent content such as pornography, computer viruses and extremist material [4]. Owing to the size and volume of spam content across the Twitter platform, attempts to perform manual classification and removal of spammers and their material are becoming increasingly futile. With this in mind, today an increasing focus of research within this domain is looking toward the adoption of machine learning techniques as a means of recognising and classifying spamming behaviours as they emerge over time.

In the work of Benevenuto et al. [4], the authors highlight how the growing emergence of spam across Twitter threatens to cause severe disruption to the platform's real-time search capabilities, by attenuating the quality of search data from which results are extracted. In particular, they focus on how Trending topics - topical content receiving the greatest attention at a given point in time - are becoming an effective new tool through which spammers are able to dramatically expand their potential audience base simply by including words associated with that given topic (typically referred to as 'hashtags') as part of their spam messages.

This paper aims to provide a discussion on the employment of machine learning techniques in combating against the proliferation of spam across the Twitter network. In particular, focus is given to the approach discussed in the work of Benevenuto et al. and their attempts to classify and subsequently detect spammers and spam material amongst real-world datasets sourced directly from Twitter.

The outline for the remainder of this paper is as follows. Section 2 introduces the work of Benevenuto et al. [4], providing discussion on their adopted approach toward classifying spammers. Section 3 proposes an extension to the work discussed in Section 2, which aims to deliver better classification accuracy through the adoption of cluster analysis on spam content attributes. Finally, Section 4 concludes the paper by discussing the limitations and potential directions for the future development of machine learning techniques to aid in the ever increasing fight against novel forms of spamming behaviour.

II. Application of Machine Learning to Spam Detection

In presenting their solution to detecting spammers operating across the Twitter platform, Benevenuto et al. decompose their approach into three core stages: dataset construction, user attribute analysis and supervised 'spammer vs. non-spammer' account classification [4]. This section provides discussion into each individual stage, with particular attention given to the relative strengths and weaknesses in the authors' approach.

A. Dataset Collection

Prior to beginning implementation of their supervised classification approach, it would first become necessary for the authors to construct a labelled dataset in which examples of both spammer and non-spammer user accounts were represented.

In the work of Benevenuto et al., the authors employ a series of web crawlers to extract user data directly from the Twitter platform [4]. Using access provided via the Twitter web API [5], these crawlers are reported to achieve 100% index coverage across the available user-base, equating to a total of 54,981,152 individual user accounts at the time of writing [4]. Any accounts setup to be private (i.e. no public access to account information) were discounted from any further analysis (approx. 8% of the total set) [4].

Looking to target the most prolific cases of spam behaviour, the authors restricted their search to target posts made in relation to any of three specially-selected Trending Topics from 2009: (a) the death of Michael Jackson; (b) the rise of Susan Boyle on talent show '*Britain's Got Talent*'; and (c) the weekly occurrence of hashtag '*#musicmonday*' [4].

To assist volunteers in their labelling efforts, the authors developed a website to facilitate the classification of an individual user based upon the content of tweets *exclusively* related to at least one of the three trending topics [4]. To minimise the risk of human error, each case was independently assessed by two volunteers. In the case of a tie situation, a third mediator was be employed to settle on a decision. The authors report a high level of general agreement across the collection of reviewed cases, which they state "*reflects a high level of confidence to this human classification process*" [4]. At the end of the labelling process, a total of 355 spammers and 7,852 non-spammers were identified. Given the significant imbalance between the proportion in user groups, this dataset was reduced to adopt a 1:2 ratio of spammers to non-spammers (710 non-spammers), leading to a final dataset size of 1,065 users [4].

B. Identification of Discriminatory User Attributes

Looking to accurately distinguish between spammers and non-spammers, Benevenuto et al. devote a significant proportion of their investigation toward the exploitation of user attributes as a means of exposing differences in their behavioural traits [4]. Under efforts to characterise behavioural traits in a formal capacity, the authors categorise their features into two distinct attribute groups, specifically: content attributes and user behaviour attributes [4].

In the case of content attributes, attention is drawn toward the composition and structure of the text contained within individual tweets. By analysing a selection of posts attributed to a particular user, the authors are able to build up a profile that captures the unique characteristics pertaining to each individual user's specific style of writing. In light of this, the authors predict that it may be possible to discriminate spammers from regular users by looking into differences observed between their associated writing styles [4]. Out of a total of 39 individual content attributes that were identified, three in particular are shown to demonstrate high discriminatory performance in relation to the detection of 'spammer-like' tendencies. From their observations made with respect to labelled spammer accounts, the authors report significant increases in cumulative frequency distribution for attributes referring to: the proportion of tweets containing URLs; the proportion of tweets containing common spam words; and the ratio of hashtags to standard words per individual tweet [4].

Benevenuto et al. also look to further support their discrimination of spammers through the capture of higher forms of behavioural disparity exhibited in the interactive conduct of users [4]. Representing these behaviours in the form of user behaviour attributes, the authors identify 23 metrics that describe the interaction and socialistic behaviours of users in the labelled dataset. These attributes are determined on a user-by-user basis, and amongst the best performing include: the ratio of followees to subscribers; the number of unique tweets that are received and the relative age of the account [4]. In all three cases, the authors report significant differences in the observations made between the spammer and non-spammer user types. For example, normal users tend to boast significantly higher numbers of subscribers than spammer accounts owing to the often failed attempts by spammers to follow large numbers of users in the hope those connections will be reciprocated. Spammer accounts also tend to exhibit much shorter lifespans than regular accounts, owing - it is assumed - to their ongoing removal performed in response to reports of spamming behaviour [4].

To substantiate the relative importance of their selected user attributes, the authors later employ two independent feature selection methods (Information Gain and Chi-Squared) to rank the attributes in terms of their discriminatory influence on the detection of spammer-operated accounts [4].

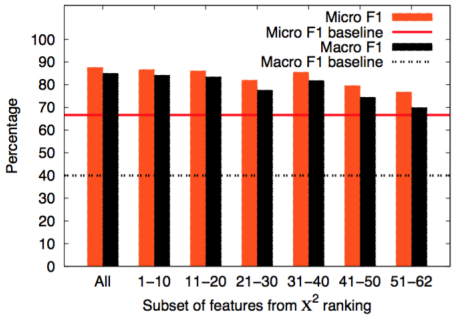

Fig. 1. Grouped rank results according to X^2 feature selection [4].

This analysis produces some interesting results, with the report indicating how in general, behaviour-led attributes appear to demonstrate better discriminatory performance over their content-led counterparts (ratio of 60:40 behaviour-led to content-led situated within the top 10 ranking results). Perhaps of most interest in amongst the findings is the demonstration of comparable discriminatory performance between that of solely the top 10 ranking attributes, and the collective of all sets (60 attributes in total) (Figure 1) [4]. This discovery leads to the determination that only a moderate number of attributes may be required in order to achieve sufficient classification performance, taking in account the assumption that overall membership of these attributes falls into the higher-ranking discriminatory sets.

C. Classification

In order to validate the discriminatory performance of their selected user attributes, the authors apply a supervised classification algorithm to a collection of unseen user accounts representative of their testing set. Given that their specific objective took the form of a binary classification problem (i.e. a user was only to be classified as either a spammer or non-spammer) the authors chose to utilise a Support Vector Machine (SVM) as the basis for their classifier [4]. To allow for cases where complex non-linear separations may exist between data points, the authors equip their SVM with a kernel function (namely the Radial Basis Function) to project the original data points into a transformed feature space that allows for standard linear separation [4] [6].

Within their approach toward classifier evaluation, Benevenuto et al. choose to adopt *K-fold cross-validation* to assess how well their supervised learning model is able to generalise to new cases of user classification. Unlike that seen for simpler approaches such as the *Holdout method*, *K-fold cross-validation* looks to reduce the level of variance in assessment scoring by averaging the performance recorded over K-runs, in which each run sees K-1 partitions of the original dataset used for training, and the one remaining partition used for validation [7]. This approach works to ensure that an evaluation does not become overly dependent on the particular makeup of either the testing or validation sets, and can also provide a more effective use of datasets in situations where labelled examples can prove difficult to acquire [8].

In the results collected from their classifier experiments, the authors report the true-positive rate for spammer classification to be 70.1%, and a true-positive rate for non-spammer classification of 96.4%. Drawing particular attention to a reported spammer misclassfication rate of nearly 30%, the authors attribute many of these cases to having exhibited "*a dual behaviour*" bearing resemblance to both spammer and non-spammer posting behaviours [4].

In order to appreciate the effects of different tradeoffs between the increased detection of spammers and the reduction in misclassification of regular users, Benevenuto et al. go on to repeat their classifier experiments under a variety of cost parameter configurations [4]. In their findings, the authors confirm their initial predictions that prescribed the existence of a common positive relationship between the value of cost parameter J, and both the rates of correct spammer classification and non-spammer misclassification. When discussing these results, the authors highlight how in the context of a real application, selecting the most appropriate value for the cost parameter will depend entirely on the individual objectives set out in relation to that specific case.

III. Expansion Proposal

One of the most pressing limitations arising from the work of Benevenuto et al. refers to the high degree of brittleness exhibited by content-led attributes when used as metrics for spammer vs. non-spammer classification [4]. This problem comes as a direct result of spammers continuously looking to undermine the existing knowledge held by supervised classification systems in support of their detection.

This paper presents a proposed expansion to the work carried out by Benevenuto et al. that seeks to improve the robustness demonstrated by supervised classification systems toward the deliberate manipulation of content-led profiling. At its core, the proposal sees the employment of unsupervised clustering techniques as a means of looking to detect and track changes in spammer writing behaviours as they adjust over time.

Cluster analysis is a widely popular form of dimensionality reduction technique [9], that aims to discover patterns in unstructured data by attempting to group together instances that are in some form more similar to one another, than to instances belonging in other disparate groups [10]. As an unsupervised learning method, clustering has the particular benefit of working with datasets that do not contain labelled examples [11]. This is vitally important when faced with situations where knowledge of expected features cannot be known in advance. With this in mind, clustering techniques have begun to witness a surge in adoption within applications looking to explore patterns exhibited in user behaviour (e.g. behavioural profiling of players active within Massively Multiplayer Online Role-Playing Games (MMORPGs) [12]).

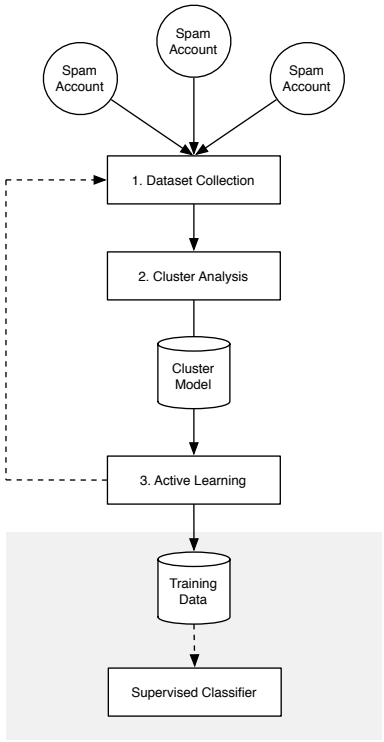

Fig. 2. Overview of three-stage process for tracking changes in spammer writing behaviours over time.

Figure 2 presents an overview of the proposed three-stage process for tracking shifts in the content-led behaviour of spammers over their operational lifetime. In the first stage, a new dataset consisting entirely of spam-related tweets is

created. Working in collaboration with the approach presented by Benevenuto et al. [4], tweet information is sourced directly from accounts previously classified as being that of a spammer. To prevent against future stagnation of dataset coverage, an approach originally inspired by [13] sees newly-classified spammer accounts automatically becoming enrolled into the growing collection of dataset sources. As the size and range of monitored spam perpetrators develops over time, one would expect to see an ongoing decline in the likelihood of encountering an entirely novel form of spam composition, arising as an inverse consequence of increasing surveillance coverage.

In the second stage of the process, cluster analysis is applied to the unlabelled dataset using the *K-means* centroid-based clustering algorithm [14]. Using an appropriate distance measure (e.g. Euclidean Distance or Cosine Distance [9]), centroid-clustering seeks to minimise the squared distance from each data point to the nearest central vector that represents an individual cluster [10]. With each of the existing content attributes representing a different dimension in the N-dimensional search space (e.g. see Figure 3), changes that are detected in the size and shape of clusters may provide an important indication as to the gradual shifts in behaviour taking place in how spammers set about composing their unsolicited content.

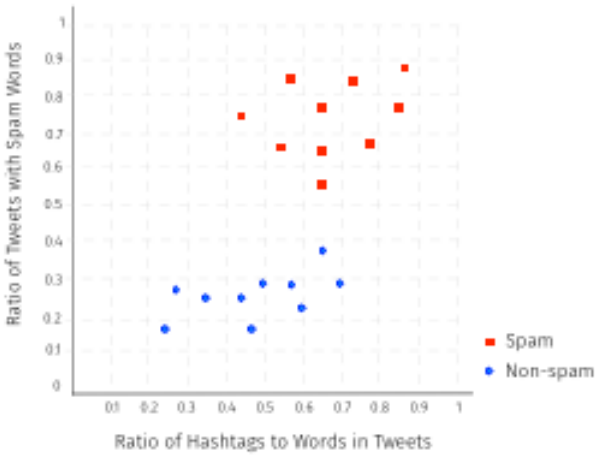

Fig. 3. Example input-space eligible for cluster analysis.

Although other types of clustering algorithms would also stand prime for consideration, such as hierarchical partitioning using a divisive model [14], centroid-based algorithms like that of *K-means* bear the particular strength of demonstrating a high level of computational performance even when faced with very large and complex datasets, as is the case here. Unlike hierarchical portioning which demonstrates a deterministic clustering strategy [6], K-means clustering uses an "*iterative improvement technique*" [15] that allows for the progressive refinement of its partitioning and clustering performance as it converges toward an optimal clustered state.

In the third and final stage of the analysis process, examples of new or significant forms of spammer behaviour, derived from the results obtained through clustering, are selected to be labelled and fed back into the training data that is then reapplied to the original supervised classification algorithm. Constituting a form of active learning, this mechanism comes inspired from a similar application proposed in [16], that incorporates clustering into its high-level methodology for continuously refining the detection accuracy of a supervised random forest classifier. Coincidentally, this paper also seeks to address the identification of spam in email messages.

Of course as with any proposal, there are limitations that arise out of selecting this particular configuration for tackling what is, at its heart, a highly complex and subjective undertaking. One significant limitation attributed to the K-means clustering algorithm is the high degree of sensitivity that it tends to exhibit toward outlier examples [17]. Given the behaviour of spammers is, by definition, an uncontrolled measure, there arises the significant possibility of outliers existing within the collected dataset. A common form of mitigation against this issue, is to instead use the *K-medians* [14] variant of the same clustering algorithm, as this supports greater robustness in the presence of outliers [17].

Another common issue with using partitioning clustering algorithms such as K-means or K-median is how to decide on the number of K clusters to use [10]. Given the ubiquity of K-means and its variants in unsupervised learning applications, many potential solutions to this issue have been proposed over time.

One of the earliest propositions to arise in this area was the *Elbow method*, which uses the Sum of Squared Error plotted against the number of clusters to try and determine the optimum point in which the SSE remains low, whilst still maintaining a sensible number of clusters (i.e. not too few clusters that the SSE measures high, and not too many clusters such that each data points effectively represents an individual cluster) [18]. Whilst in theory this approach shows great promise, in reality it can prove highly difficult to specifically locate this optimum value to determine the number of clusters [17]. Another approach sees the employment of cross-validation to test multiple values for K over a series of tests and training/test datasets [19]. In this case, the value for K shown to minimise errors in the test set typically gets selected.

Owing to constant efforts by service providers to block and disable spam-related activities, it is commonplace for accounts linked to these sorts of activity to remain active for much shorter amounts of time than those operated by regular users [4]. In the context of this proposal, this has the unfortunate consequence of dramatically reducing the ability to monitor *individual* accounts over a longer time-span that may help to uncover more gradual changes in spammer writing behaviours. Although this issue gives initial cause for concern, it may not be as first thought, given that these trends are likely to remain visible across the wider population of spammer accounts and are not only isolated to individual cases.

IV. Future Work & Conclusion

Techniques focused on the detection and removal of spammers are becoming evermore sophisticated thanks to the significant contributions made in the field of machine learning and data analysis. Unfortunately despite the best efforts of research teams and service providers alike, spammers continue to employ equally creative measures to manipulate and undermine the exact systems put in place to tackle them. This paper identifies two areas in which future research focus is likely to lead toward significant improvements in the accuracy and robustness of systems aiming to purge online platforms of spammers and their content.

The first area seeks greater use of URL analysis as a primary identifier of spam-related content. Although the vast majority of URL links posted by spammers obfuscate their real location through URL shortening services, new techniques are now emerging that show impressive resilience to these tactics through the analysis of click-through rates between cases of genuine and spam URLs [20].

The second area looks toward the continued refinement of supervised classification approaches through the application of active learning in the generation of labelled training sets. This approach has already seen successful implementation in the domain of spam detection bearing impressive performance over manual labelling techniques [21], however this paper finds no current evidence of the same approach being applied to the purpose of isolating spammer accounts as well as spam content.

With an ongoing increase expected in the demand and popularisation of social media, platforms like Twitter, Facebook and Instagram will continue to represent high priority targets for spammers in the coming years. Whilst it is unfeasible to envisage a time where spam has completely been eradicated from the digital landscape, the approaches highlighted in this paper present a highly compelling case for the future development of advanced machine learning techniques capable of promising a significant reduction in the day-to-day impact of spam upon the online social endeavours of regular users.

References

[1] A. Perrin, "Social Media Usage: 2005-2015," *Pew Research Center*, 2015.

[2] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden, "Social Media Update 2014," *Pew Research Center*, vol. 19, 2015.

[3] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, Jul. 2010, pp. 1–10.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, 2010, p. 12.

[5] "Twitter Developers - REST APIs," https://dev.twitter.com/rest/, 2016.

[6] C. D. Manning, P. Raghavan, H. Schütze, and Others, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[7] C. Elkan, "Evaluating classifiers," *University of San Diego, California, retrieved [01-11-2012] from http://cseweb. ucsd. edu/˜ elkan B*, vol. 250, 2012.

[8] J. Schneider, "Cross Validation," https://www.cs.cmu.edu/~{}schneide/tut5/node42.html, Feb. 1997.

[9] A. Drachen, "Introducing Clustering: Behavioral Profiling for Game Analytics," http://andersdrachen.com/2014/05/13/introducing-clustering-i-behavioral-profiling-for-game-analytics/, May 2014.

[10] C. Lu, "SEM6420: Clustering," Aberystwyth University, Feb. 2016.

[11] "Unsupervised Learning: Machine Learning with MATLAB," http://uk.mathworks.com/discovery/unsupervised-learning.html, Mathworks Inc, 2016.

[12] A. Drachen, R. Sifa, C. Bauckhage, and C. Thurau, "Guns, swords and data: Clustering of player behavior in computer games in the wild," in *Computational Intelligence and Games (CIG), 2012 IEEE Conference on*, Sep. 2012, pp. 163–170. [Online]. Available: http://dx.doi.org/10.1109/CIG.2012.6374152

[13] K. Lee, J. Caverlee, and S. Webb, "Uncovering Social Spammers: Social Honeypots + Machine Learning," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 435–442.

[14] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[15] A. LaPaugh, "Clustering Algorithms: Divisive hierarchical and flat," Princeton University, Feb. 2011.

[16] D. DeBarr and H. Wechsler, "Spam detection using clustering, random forests, and active learning," in *Sixth Conference on Email and Anti-Spam. Mountain View, California*. Citeseer, 2009.

[17] P. Rai, "Data Clustering: K-means and Hierarchical Clustering," University of Utah, Oct. 2011.

[18] R. Gove, "Using the elbow method to determine the optimal number of clusters for k-means clustering," https://bl.ocks.org/rpgove/0060ff3b656618e9136b, Dec. 2015.

[19] "Finding the Right Number of Clusters in k-Means and EM Clustering: v-Fold Cross-Validation," http://www.statsoft.com/Textbook/Cluster-Analysis#vfold, 2016.

[20] D. Wang, S. B. Navathe, L. Liu, D. Irani, A. Tamersoy, and C. Pu, "Click traffic analysis of short URL spam on Twitter," in *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on*, Oct. 2013, pp. 250–259.

[21] C. Thongsuk, C. Haruechaiyasak, and P. Meesad, "Classifying Business Types from Twitter Posts Using Active Learning." in *IICS*. Citeseer, 2010, pp. 180–189.