Procedia Computer Science

Volume 53, 2015, Pages 224–231

2015 INNS Conference on Big Data

Automated detection of human users in Twitter

M.A. Fernandes, P. Patel, and T. Marwala

Department of Electrical and Electronic Engineering University of Johannesburg,Johannesburg,
Gauteng, South Africa

Abstract

This paper compares Suppport Vector Machine (SVM) classification and a number of clustering approaches to separate human from not human users in Twitter in order to identify normal human activity. These approaches have similar F_1 accuracy scores of 90% with both experiencing difficulties in classifying human users behaving abnormally. A second stage classification step was then used to further separate not human users into brands, celebrities and promoters / information achieving an average F_1 accuracy of 74%. These accuracies were achieved by reducing the size of the feature space using stepwise feature selection and category balancing from manual inspection of classification results.

Keywords: Twitter, SVM, Clustering, Human

1 Introduction

Microblogging, and Twitter in particular, has recently attracted a large number of users. Twitter has roughly 5.5 million active users in South Africa (284 million globally) tweeting over 5 million times a month. Twitter is a platform where people can share their thoughts and ideas (often anonymously) with friends or the population as a whole using 140 characters or less. Twitter is a good choice for initial research studies due to its open nature [5].

Due to the different uses of and agendas on the platform it can often be difficult to identify users that can be mined for meaningful information. Previous work on removing noise generated from other user types from Twitter data has focussed on detecting a subset of users called spammers. Spammers are a type of user who contaminate the information exposed to other, legitimate, users for their own (often malicious) purposes which can lead in turn to a risk to the security and privacy of social networks. Spammers can be seen to belong to one of the following categories [15]:

1. **Phishers.** These users behave like normal users in order to obtain personal and sensitive information belonging to other users.

2. **Fake users.** These users pretend to be someone they are not by impersonating someone else's account to send spam content to their network.

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015

doi:10.1016/j.procs.2015.07.298

3. **Promoters.** These users send unsolicited links and advertisements or other promotional material to other Twitter users. This is often used to obtain user personal information.

For the purpose of this study it was necessary to look beyond just identifying spammers and focus on excluding non-human users or accounts as they would not be amenable to targeted content. As such this paper proposes the application of a combination of supervised and unsupervised techniques to isolate these abnormal individuals from the broader population of South African Twitter users. A number of behavioural metrics were created based on the metadata collected from 14 month of tweets and used as inputs in the classification and clustering processes.

The rest of this paper is organised as follows: a brief overview of related work, a description of the methodology applied, details of the experiments and associated results and ending with conclusions derived.

2 Related Work

Identification of anomalous user types in Twitter data is an important precursor to detailed analyses of Twitter behaviours as they could incorrectly skew the results obtained in terms of topics prevalent in the population. Identification of specific types of users as different from the rest of the population is, in essence, a form of creating a profile of the user's interactions with the platform.

Existing techniques in spammer detection typically use a pre-classified data set and a combination of behavioural (content, user information, network and topic) to create a classifier that can accurately differentiate spammers from legitimate users with accuracies obtained of around 90%. The main difference in the majority of these approaches is in the features used for classification [16, 1, 3, 12].

Chakraborty et. al. [4] proposed a slightly different system to detect users posting abusive content such as harmful URLs, porn URLs, and phishing links as part of the friend request process. The solution was tested on 5 000 accounts with the SVM classifier performing the best, achieving an accuracy of 89%.

Miller et. al. [13] attempt to treat the identification of spammers as an anomaly detection and not classification problem where outliers are flagged as spammers. They utilise a combination of user metrics and one gram text features. They then test two algorithms: DBSCAN which uses a density based similarity metric and K-Means which uses an Euclidean distance based metric. These approaches achieved an 82% and 71% F_1 score respectively with high accuracy but low precision.

3 Methodology

The approach followed in this paper was to define types of users that would be of interest. These definitions were then used to manually classify a sample of users as input into a classification algorithm with a number of metrics measuring behavioural characteristics. This feature space was then decreased using feature selection and applied to classification and clustering algorithms for comparison purposes.

3.1 User Type Definitions

Through investigation of twitter users the following main user types were defined:

Table 1: Features created from Twitter information.

Feature Type	Feature
Content	URL usage, Hashtag usage, Retweets, Mentions, Swear words, Exclamation marks, Abbreviations, Emoticons
User	Source type, Number of sources, Reputation, Friends, Followers, Active days, Tweet rate, Time of day, Day of week
Network	Mentioned, Unique mentions
Topic	Popular word count

1. **Person.** This user type posts information on their own behalf in order to chat with friends, engage on a topic or for a variety of other reasons.

2. **Promoter.** This user type consistently broadcasts unsolicited advertisements or offers for a specific good or service.

3. **Celebrity.** This user type is a well-known world personage with a fan base that is not just in the social media space who can be very conscious of their public relations.

4. **Brand.**This user type is the official mouthpiece for an organisation. The account is used to interact with their customers, dealing with service issues and promoting their services.

5. **Information.** This user broadcasts news and information and content, not necessarily to a large number of followers.

Manual classification was performed by looking at every tweet submitted by and to the user and deciding which category would best fit. The user category was determined by searching for name matches with celebrities and brands or looking at the topics of the tweets.

3.2 Feature Extraction

Based on previous studies [15] it was decided to create a combination of content, user, network and topic features as summarised in Table 1. These features were derived from tweet information as well as other metadata available via the Twitter API and in total added up to 70 variables.

3.3 Feature Selection

The feature set was initially reduced by removing highly correlated variables as these contain the same information biasing the analysis accordingly [9]. The optimum number of features was then determined by backward stepwise selection where features are recursively removed from the population, a new classifier is trained and the cross validated accuracy tested. At each step the feature with the lowest coefficient weights was removed. The outputs of this recursion was then analysed to see what the optimum set of features was [2][8].

3.4 Classification

This paper investigated the use of a two stage classification model that initially separated human and not human and then separated not human into either celebrity, brand or promoter / Information. The reason for grouping the last two categories was that they were difficult to separate due to similarities in behaviour.

The classification algorithm used was Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel. SVMs, in particular the LIBSVM implementation [10], were selected as they are a popular classification algorithm whose performance is easy to understand and compare to other methods. Other kernel functions such as linear and polynomial were tested but their performance was seen to be inferior as compared to RBF and so were not included in this paper [6].

SVMs function by mapping data into a new feature space so that a linear classifier can be used. The mapping is done with the use of kernel functions and is necessary because it may be impossible to have a classifier that divides the data in linear space. The two main focuses of SVM algorithms are therefore to find a kernel function that optimally divides the data and to find the maximal margin hyperplane [6].

3.5 Clustering

This paper investigated the use of different clustering algorithms to separate Twitter users into human and not human. Clustering was attempted as an automated mechanism for performing user separation without pre-labelled data. The following three popular clustering algorithms were assessed: K-Means, Gaussian Mixture Models and DBSCAN [7]. The scikit learn implementation [11] was used as the basis for each.

The K-Means algorithm partitions data sets into K distinct, non-overlapping clusters. K points are randomly generated which serve as cluster centroids and then assigns each data point to a cluster by calculating the Euclidean distance to the cluster centres and identifying the closest centroid . The algorithm then iteratively adjusts the centroids based on the average of all clustered points and re-assigns the points to the closest cluster until stable [7].

A common method to determine the optimal value of K to be used in the K-Means algorithm is the Elbow Method. This quantifies the percentage of variance explained as a function of the number of clusters such that the rate of information gained (explained variance) by adding another cluster drops thereby creating an angle or elbow in the graph. Percentage of variance explained can be defined as the ratio of the between-group to the total variance [7].

A Gaussian mixture model (GMM) is conventionally defined as a probabilistic model that assumes all data points in the population are generated from a mixture of a finite number of Gaussian distributions whose unknown parameters can be estimated. These distributions are conventionally regarded as a generalisation of the K-means algorithm incorporating information about the covariance structure of the data as well as the centres of the latent Gaussians. The algorithm uses expectation maximisation to determine the maximum likelihood of the latent parameters [7].

The DBSCAN algorithm separates data according to density with clusters being areas of high density separated by areas of low density. As a result of this density based approach the clusters identified by DBSCAN can be any shape, as opposed to K-Means which only identifies convex shaped clusters. A cluster is a set of core samples which are close to each other in terms of distance and a set of non-core samples that are close to a core sample. A cluster needs to satisfy the following properties: all points are mutually density connected and every point is density reachable from anywhere in the cluster [7].

The performance of these clustering algorithms was then calculated in a manner comparable to the SVM classification using the knowledge of the ground truth. The clusters were translated to a category classification based on the dominant category of the cluster. The more homogenous the clusters the more accurate the algorithm was in terms of classified category.

3.6 Performance Evaluation

The performance of the algorithms was determined using cross validated accuracy scores. Accuracy was determined from the confusion matrix created from the classification and misclassification of each category. The F_1 score is the harmonic mean of the precision (True positives / (True positives + False positives))and recall (True positives / (True positives + False negatives))[14].

4 Experiment

4.1 Data Preparation

The data was filtered to users with more than 10 tweets over the period as well as English language tweets to limit analysis to users where there was sufficient information to make a judgment and to tweets that could be analysed for English content. A random sample of 1 000 users was selected to be manually classified. This was further enhanced through building a classifier on the sample and then investigating and manually classifying users classified as not human, thereby balacing the data set.

The calculated features were transformed so that the distributions were uniform and differences were comparable. The main transformations performed were changing variables to a log scale and then scaling the range to fall between zero and one.

4.2 Feature Selection

Feature selection was performed separately for each stage of classification: human vs not and brand vs celebrity vs promoter / information. This was done as the behavioural characteristics between these two tasks were very different.

The number of features selected for stage one and the clustering exercise was twenty with an expected F_1 score of 90%. Table **??** shows which features were selected as a result of this process. The number of features selected for stage two was thirty eight with an expected F_1 score of 70%. Table **??** shows which features were selected as a result of this process. Stage one predominantly used features relating to followers, content, some source types and many key words. Stage two, on the other hand, used more hashtags, sources, days and time of day.

4.3 Classification Model

A grid search was performed to identify the optimum values for the classification parameters (C and gamma) where these values were incremented using a log scale. The results for stage one are shown in Table 2 with **C = 128** and **gamma = 0.03125**. These parameters were then used to train a classifier using cross fold validated sections of the training data. The best solution was then used on a held out data set to measure the accuracy of the classification model, the confusion matrix from this testing data can be seen in Table 2. The classifier had difficulties in correctly labelling humans that did not behave normally.

The F_1 accuracy scores for the stage one classifier can be seen in Table 3. Although the scores were very similar it can be seen that the model created was slightly more accurate in identifying human users. This result allows for further, more focused, analysis performed on human users. Support is the number of cases or users belonging to the class.

A grid search was performed to identify the optimum values for the classification parameters (C and gamma). The results for stage two are shown in Table 4 with **C = 128** and **gamma = 0.03125**. These parameters were then used to train a classifier on cross fold validated sections

Table 2: Confusion matrix for stage one classifier.

Category	Predicted Human	Predicted Not Human
Human	294	36
Not Human	14	204

Table 3: Accuracy scores for stage one classifier.

Category	Precision	Recall	F_1	Support
Human	0.95	0.89	0.92	330
Not Human	0.85	0.94	0.89	218
Average	0.91	0.91	0.91	548

of the training data. The best solution was then tested on a held out data set to measure the accuracy of the model. The confusion matrix from this testing data can be seen in Table 4.

The accuracy scores for the stage two classifier can be seen in Table 5. The classifier was good at identifying Promoters or Information user with an 80% F_1 score. It, however, could not differentiate between these two, hence the two categories were grouped. The accuracy on Celebrities and Brands was somewhat lower due to the variation and similarities in behaviour across these two categories.

4.4 Clustering Models

In order to identify the optimal number of clusters inherent in the data an elbow analysis of the information gain was performed. Although the interpretation of these results can often be subjective there was a clear elbow in the data at around two to three clusters. As detailed in the methodology section these clusters where then converted to classifyers using the homogeneity of the cluster in terms of human and not human. Table 6 shows the confusion matrix of the results of this process.

The accuracy scores for the different clustering algorithms can be seen in Table 7. The best performing clustering algorithm was K-Means but the other two still performed well with over 80% F_1 score. All of the clustering algorithms have a similar issue with people who do not behave normally in that they have a high ratio of URLs and mention a lot of different users.

4.5 Discussion of Results

The classification and clustering algorithms both achieved a 90% F_1 score on separating humans from not humans and had similar issues with humans behaving abnormally. Incorporating

Table 4: Confusion matrix for stage two classifier.

Category	Predicted Celebrity	Predicted Brand	Predicted Information / Promoter
Celebrity	36	4	0
Brand	15	41	8
Information / Promoter	6	9	45

Table 5: Accuracy scores for stage two classifier.

Category	Precision	Recall	F_1	Support
Celebrity	0.63	0.90	0.74	40
Brand	0.76	0.64	0.69	64
Information / Promoter	0.85	0.75	0.80	60
Average	0.76	0.74	0.74	164

Table 6: Confusion matrices for clustering (1=GMM, 2=K-Means, 3=DBSCAN).

Category	Predicted Human	Predicted Not Human
Human	726(1), 841(2), 733(3)	236(1), 159(2), 267(3)
Not Human	15(1), 29(2), 12(3)	644(1), 630(2), 647(3)

additional topic information would increase the accuracy of classification as these groups would talk about different things. The accuracy on the clustering algorithms was achieved through feature reduction. This approach could be further enhanced by looking at reducing features without incorporating any apriori knowledge.

The stage two classifier performed well in identifying celebrities but tended to include some brands in this class. The Information / Promoter users were well separated but sometimes misclassified as the other categories. These overlaps were due to the fact that some users could be exhibiting multiple behaviours, for example a brand trying to promote their products.

5 Conclusions

This paper compared classification and clustering approaches to separate human from not human users in Twitter. An initial feature set of 70 variables was reduced to the most relevant for classification, thereby decreasing complexity and improving generalisation performance.

The classification and clustering approaches had similar F_1 accuracy scores of 90% and experienced difficulties in classifying human users who behaved abnormally. A second stage classification step was used to separate not human users into brands, celebrities and promoters / information achieving an average F_1 accuracy of 74%. The classification and clustering approaches perform with equal accuracy in separating human and not human users but clustering has the advantage of not requiring pre-classified data.

Table 7: Accuracy scores for clustering algorithms.

Category	Algorithm	Precision	Recall	F_1
Human	GMM	0.98	0.76	0.86
	K-Means	0.97	0.84	0.90
	DBSCAN	0.98	0.73	0.84
Not Human	GMM	0.73	0.98	0.84
	K-Means	0.80	0.96	0.87
	DBSCAN	0.87	0.83	0.83

The categories were balanced through recursively running the model on the full population and investigating the results. This process reduced the influence of the dominant human category and hence allowed for the creation of more stable classifiers. Further repetitions would continue to refine the models. In addition to this more detailed topic or content features would assist in improving the accuracies.

References

[1] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang. Cats: Characterizing automation of twitter spammers. Technical report, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, 2013.

[2] H. An, D. Huang, Q. Yao, and C. Zhang. Stepwise searching for feature variables in high-dimensional linear regression. Technical report, Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing 100080, China.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, Redmond, Washington, US, 2010.

[4] A. Chakraborty, J. Sundi, and S. Satapathy. Spam: A framework for social profile abuse monitoring. Technical report, Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-4400, USA, 2012.

[5] Blue Magnet Digital. The State of Social Media in South Africa in 2013. `http://www.bluemagnet.co.za/blog/the-state-of-social-media-in-south-africa-2014`, 2013. [Online; accessed 30-May-2015].

[6] P. Erasto. *Support Vector Machines - Backgrounds and Practice*. PhD thesis, Rolf Nevanlinna Institute, 2001.

[7] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. SIAM, Society for Industrial and Applied Mathematics, 2007.

[8] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research 3 (2003) 1157-1182*, 2003.

[9] M. A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Department of Computer Science, University of Waikato, 1999.

[10] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2010.

[11] Scikit Learn. Clustering. `http://scikit-learn.org/stable/modules/clustering.html`, 2014. [Online; accessed 30-May-2015].

[12] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In Jose M. Alcraz Calero, editor, *Autonomic and Trusted Computing: 8th International Conference, Atc 2011, Banff, Canada, September 2-4, 2011, Proceedings*, Bannf, Canada, 2011. Springer.

[13] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. Twitter spammer detection using data stream clustering. Technical report, Department of Computer Science, Houghton College, Houghton, NY, USA, 2014.

[14] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. Technical report, SITE, University of Ottawa, 2006.

[15] M. Verma, Divya, and S. Sofat. Techniques to detect spammers in twitter - a survey. *International Journal of Computer Applications (0975 - 8887)*, 85(10):27–32, 2014.

[16] A. H. Wang. Don't follow me - spam detection in twitter. Technical report, College of Information Sciences and Technology, The Pensylvania State University, PA 18512, Dunmore, USA.