Content Feature Extraction in the Context of Social Media Behavior

Shai Neumann[1], Charles Li[2], Chloe Lo[3], Corinne Lee[3],
Shakeel Rajwani[3](✉), Suraj Sood[3], Buttons A. Foster[3], Toni Hadgis[3],
Yaniv Savir[3], Frankie Michaels[3], Alexis-Walid Ahmed[3],
Nikki Bernobic[3], and Markus Hollander[3]

[1] Eastern Florida State College, Melbourne, FL, USA
neumanns@easternflorida.edu
[2] Mercy College, Dobbs Ferry, NY, USA
[3] Sirius17, Melbourne, FL, USA
shakeel.rajwani@gmail.com

Abstract. Twitter accounts are used for a multitude of reasons, including social, commercial, political, religious, and ideological purposes. The wide variety of activities on Twitter may be automated or non-automated. Any serious attempt to explore the nature of the vast amount of information being broadcast over such a medium may depend on identifying a potentially useful set of content features hidden within the data. This paper proposes a set of content features that may be promising in efforts to categorize social media activities, with the goal of creating predictive models that will classify or estimate the probabilities of automated behavior given certain account content history. Suggestions for future work are offered.

Keywords: Twitter · Social media · Content feature extraction

1 Background

1.1 Introduction

Social media activity data, in the case of this paper Twitter account activity, can be understood as consisting of two primary components, metadata or demographics, and content data. Metadata involves external characteristics such as time of activity, time of account creation, location, type of platform used for activity, number of friends, followers, and more. Content data involves syntactic and semantic characteristics. The focus of this paper is on content data, in particular, content feature extraction that can be implemented on a large set of text data in order to enable categorization of types of activities and classification of activities as automated versus non-automated.

1.2 The Content Data Elements and Their Encoding

Below are some linguistic features that can be extracted from the text content generated by Twitter users. These features can be used to generate mathematical "signatures" for

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2017, Part I, LNAI 10284, pp. 558–570, 2017.
DOI: 10.1007/978-3-319-58628-1_42

different types of online behaviors. In this way, they augment account demographic features to create a rich, high-fidelity information space for behavior mining and modeling.

1. *The relative size and diversity of the account vocabulary*
   Content generated by automated means tends to reuse complex terms, while naturally generated content has a more varied vocabulary, and terms reused are generally simpler.
2. *The word length mean and variance*
   Naturally generated content tends to use shorter but more varied language than automatically generated content.
3. *The presence/percentage of chat-speak*
   Casual, social users often employ simple, easy to generate graphical icons, called emoticons. Sophisticated, non-social users tend to avoid these unsophisticated graphical icons.
4. *The presence and frequency of hashtags*
   Hashtags are essentially topic words. Several hashtags taken together amount to a tweet "gist". A table of these could be used for automated topic/content identification and categorization.
5. *The number of misspelled words*
   It is assumed that sophisticated content generators, such as major retailers, will have a very low incidence of misspellings relative to casual users who are typing on a small device like a phone or tablet.
6. *The presence of vulgarity*
   Major retailers are assumed to be unlikely to embed vulgarity in their content.
7. *The use of hot-button words and phrases ("act now", "enter to win", etc.)*
   Marketing "code words" are regularly used to communicate complex ideas to potential customers in just a few words. Such phrases are useful precisely because they are hackneyed.
8. *The use of words rarely used by other accounts (e.g., tf-idf scores)*
   Marketing campaigns often create words around their products. These created words occur nowhere else, and so will have high tf-idf scores, which is the term frequency–inverse document frequency score.
9. *The presence of URL's*
   To make a direct sale through a tweet, the customer must be engaged and directed to a location where a sale can be made. This is most easily accomplished by supplying a URL. URL's, even tiny URL's, can be automatically followed to facilitate screen scraping for identification/characterization.
10. *The generation of redundant content (same tweets repeated multiple times)*
    It is costly and difficult to generate unique content for each of thousands of online recipients. Therefore, automated content (e.g., advertising) tends to have a relatively small number of stylized units of content that they use over and over. The result is an account with "redundant" content.

2 Method

2.1 Data

Twitter account activity data is available through the Twitter API (application program interface) which returns requests for random samples of data in the JSON (JavaScript Object Notation) data structure containing both demographics and content.

Content data (tweets) are returned (in the JSON structure) as character strings of length 1 to 140 characters. They may be in any language or no language at all. Tweets can contain any combination of free text, emoticons, chat-speak, hashtags, and URL's. Twitter does not filter tweets for content (e.g., vulgarisms, hate speech).

For this study a sample of the activities of 8845 Twitter accounts containing the content of 1,048,395 tweets was collected for content analysis.

2.2 Procedures

A vector of text features is derived for each user. This is accomplished by deriving text features for each of the user's tweets and then rolling them up, i.e. summing and normalizing the data. Therefore, one content feature vector is derived for each user from all of that user's tweets.

The extraction of numeric features from text is a multi-step process:

1. Collect the user's most recent (up to 200) tweet strings into a single set (a Thread).
2. Convert the thread text to upper case for term matching.
3. Scan the thread for the presence of emoticons, chat-speak, hashtags, URL's, and vulgarisms, setting bits to indicate the presence/absence of each of these text artifacts.
4. Remove special characters from the thread to facilitate term matching.
5. Create a Redundancy Score for the Thread. This is done by computing and rolling up (sum and normalize) the pairwise similarities of the tweet strings within the thread using six metrics: Euclidean Distance, RMS-Distance, L1 Distance, L-Infinity Distance, Cosine Distance, and the norm-weighted average of the five distances.
6. The thread text feature vector then contains as vector components user scores based on features such as the emoticon flag, the chat-speak flag, the hashtag flag, the URL flag, the vulgarity flag, and the Redundancy score.

A list of 23 potential content related features was created and calculated for each of the 8845 Twitter accounts in the sample (Tables 1 and 2).

For the purpose of classifying accounts as automated (bots) versus non-automated, a manual rating process of a sample of tweet content coming from 101 active accounts was executed. The sample was divided into 5 subsets with each set being rated by multiple volunteers who read the content of approximately 20 accounts in each subset,

Table 1. Sample of raw data

Feature		Set 1	Set 2	Set 3
UserID		22821737	22822092	22823578
1	tweets	10	190	133
2	adj	1.7	2.247368	1.774436
3	adv	0	0.2684211	0.09774436
4	art	0.1	1.994737	1.338346
5	commnoun	4.2	1.215789	1.736842
6	conj	0.6	0.6947368	0.3458647
7	interj	0	0.005263158	0.007518797
8	prep	0.6	0.3736842	0.3383459
9	pron	0	0.368421	0.03759398
10	Propnoun	1.4	1.931579	1.699248
11	verb	0.4	1.215789	0.6315789
12	stopword	0	0.06842105	0.04511278
13	vulgar	0	0.01578947	0
14	hash	0.6	0.4894737	0.1052632
15	urls	1	0.1473684	0.9774436
16	case	0	0	0
17	punc	1	0.9842106	1
18	emo_chat	0	0	0
19	good_len	82.2	74.14211	70.9624
20	good_cnt	13.3	16.08947	12.59398
21	bad_len	0.7	1.394737	1.233083
22	bad_cnt	0.1	0.2	0.1954887
23	redund	0.7686407	0.7453661	0.740773

Table 2. The list of 23 features for analysis

Feature		Description
1	tweets	Number of tweets up to 200
2	adj	Number of adjectives per tweet
3	adv	Number of adverbs per tweet
4	art	Number of articles per tweet
5	commnoun	Number of common nouns per tweet
6	conj	Number of conjunctions per tweet
7	interj	Number of interjections per tweet
8	prep	Number of prepositions per tweet
9	pron	Number of pronouns per tweet
10	Propnoun	Number of proper nouns per tweet

(*continued*)

Table 2. (*continued*)

Feature		Description
11	verb	Number of verbs per tweet
12	stopword	Number of stop words matching a list- per tweet
13	vulgar	Number of vulgar words matching a list- per tweet
14	hash	Number of hashtags per tweet
15	urls	Number of urls per tweet
16	case	Relative frequency of usage of both lower and upper case
17	punc	Relative frequency of usage of punctuation
18	emo_chat	Number of emoticons per tweet
19	good_len	Number of *characters* in correctly spelled words per tweet
20	good_cnt	Number of *words* of correctly spelled words per tweet
21	bad_len	Number of *characters* of incorrectly spelled words per tweet
22	bad_cnt	Number of *words* of incorrectly spelled words per tweet
23	redund	Redundancy Score for the Thread

each subset containing a few thousand tweets. The rating of each account involved classification as a bot or not and also the assignment of a level of confidence associated with such classification, then a brief explanation of the main reasons was given for the relevant decisions. Of the 101 accounts, 65 were classified as 35 bot accounts and 30 non-bot accounts with a high level of confidence. Those 65 accounts were then assigned a dependent variable value of 1 if identified as a bot, and 0 otherwise.

3 Results

Excel was used to generate a correlation matrix for the 23 content features for the large sample of 8845 feature vectors (Table 3).

Similarly, correlations between the 23 content features and the dependent variable for the small set of 65 accounts were calculated and sorted based on absolute value (Table 5).

Absolute values of the correlations between features and the dependent variable ranged from 0.003 to 0.603. Ranking such absolute values of correlations resulted in the following list of top predictors of bot-like behavior: “redund”, “urls”, “good_len”, “adj”, “tweets”, “vulgar”, “good_cnt”, “commnoun”, “emo_chat” and “art”.

Charts were created to examine the distributions of features that were deemed to be significant in terms of their correlation with the dependent variable in the small sample. Charts were created to examine joint distributions. Following some interpretation of the nature of distributions, some hypotheses were made as to potential statistical learning tools that may be useful in modeling based on such content features (Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11).

Table 3. Correlation among the 23 features of tweet data (correlation scores above 0.6 are bolded)

	1	2	3	4	5	6	7	8	9	10	11
1	**1.000**										
2	0.029	**1.000**									
3	-0.019	0.044	**1.000**								
4	0.032	0.110	0.436	**1.000**							
5	0.094	0.086	0.135	0.292	**1.000**						
6	0.019	0.076	0.407	**0.630**	0.214	**1.000**					
7	-0.041	-0.066	0.031	-0.104	0.104	-0.079	**1.000**				
8	0.040	0.088	0.144	0.417	0.267	0.321	-0.113	**1.000**			
9	-0.078	0.070	0.400	0.339	0.090	0.387	0.043	0.128	**1.000**		
10	0.039	0.054	0.302	0.545	0.533	0.431	0.078	0.322	0.245	**1.000**	
11	0.006	0.115	0.424	**0.701**	0.279	0.544	-0.134	0.381	0.360	0.448	**1.000**
12	-0.007	0.069	0.216	0.263	0.076	0.262	-0.063	0.179	0.285	0.152	0.277
13	-0.052	-0.014	0.072	0.038	-0.031	0.038	0.037	-0.059	0.120	-0.020	0.059
14	-0.010	-0.021	-0.028	0.021	0.119	-0.054	-0.013	0.077	-0.072	0.061	0.068
15	0.299	-0.066	-0.254	-0.216	0.028	-0.257	-0.106	0.059	-0.296	-0.147	-0.199
16	-0.149	0.190	-0.022	-0.093	-0.134	-0.026	-0.001	-0.091	0.010	-0.144	-0.070
17	0.207	-0.009	-0.034	0.123	0.156	0.053	-0.068	0.146	-0.100	0.148	0.069
18	-0.044	0.127	0.011	0.096	-0.006	-0.014	0.014	0.048	0.053	-0.002	0.123
19	0.160	0.101	0.216	0.490	0.590	0.326	-0.026	0.470	0.088	0.580	0.473
20	0.081	0.298	0.390	**0.702**	**0.650**	0.538	0.023	0.502	0.309	**0.752**	**0.665**
21	-0.047	-0.177	-0.131	-0.280	-0.170	-0.183	0.054	-0.134	-0.110	-0.220	-0.254
22	-0.035	-0.172	-0.068	-0.255	-0.105	-0.136	0.079	-0.101	-0.091	-0.166	-0.237
23	0.352	0.178	-0.015	-0.001	0.073	0.011	0.018	0.021	0.001	0.061	-0.027

4 Discussion

4.1 Findings

Approximately 10% of the 8845 accounts had the maximum level of activity measured (200 tweets). This may provide some lower bound estimate of the rate of accounts exhibiting bot-like behavior.

Examination of the content features correlation matrix reveals that correlations are generally low with some explainable exceptions. Features such as good_len and good_cnt refer to the number of characters that are part of correctly spelled words and the number of correctly spelled words, respectively. The high correlation of 0.86 is to be expected, and such is the case for bad_len and bad_cnt with a correlation of 0.841

Table 4. Correlation among the 23 features of tweet data

	12	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
12	**1.000**											
13	0.009	**1.000**										
14	-0.014	-0.054	**1.000**									
15	-0.048	-0.225	0.088	**1.000**								
16	-0.018	0.038	-0.136	-0.264	**1.000**							
17	0.018	-0.158	0.118	0.567	-0.350	**1.000**						
18	0.004	0.025	0.019	-0.037	-0.010	-0.021	**1.000**					
19	0.189	-0.131	0.380	0.313	-0.277	0.433	0.042	**1.000**				
20	0.266	-0.038	0.208	-0.045	-0.159	0.271	0.078	**0.861**	**1.000**			
21	-0.117	-0.029	0.086	0.104	-0.017	0.055	-0.008	-0.102	-0.211	**1.000**		
22	-0.087	-0.030	0.109	0.064	-0.008	0.057	-0.020	-0.009	-0.112	**0.841**	**1.000**	
23	0.027	-0.052	0.007	0.159	-0.187	0.145	0.007	0.103	0.098	-0.007	-0.039	**1.000**

(both highlighted in Table 4). In both situations, consideration may be given to selecting only one of each pair for the purpose of predictive modeling.

The top ten content features appear to contain discriminating information that may be relevant in an attempt to classify Twitter accounts as bot or non-bot accounts. Separation issues and the skewed nature of the majority of the distributions of content features may justify an expectation that a nonparametric approach may perform better than a parametric one.

The distribution of the redundancy scores appears to be approximately normal, while all other distributions examined are skewed. As in the case of an earlier study of external features, most relevant distributions that quantify social media behaviors do not appear to be normal, a fact that may later support preference for nonparametric modeling techniques or the application of some feature transformations.

Examination of the scatter plots of joint distributions seems to support the selection of the top content features listed above. One can note that in the case of vulgarity score

Table 5. Correlation of the 23 features to the dependent variable (bot or not Boolean value)

Feature		r score
23	redund	0.602903665143099
15	urls	0.552239841627008
19	good_len	0.499866059699615
2	adj	0.439996556749289
1	tweets	0.405312199707016
13	vulgar	-0.386187081404597
20	good_cnt	0.361167846205383
5	commnoun	0.336302040152226
18	emo_chat	-0.322361395640107
4	art	-0.306464242615507
6	conj	-0.266514973936451
12	stopword	-0.256512790006307
9	pron	-0.23235623235559
17	punc	0.22984473910942
8	prep	0.217071031951804
10	Propnoun	0.215136062319311
7	interj	-0.202111817921263
14	hash	0.125290858127832
3	adv	-0.0933858445685339
16	case	-0.0477397194562674
21	bad_len	0.0373329649121563
22	bad_cnt	0.0035443689757518
11	verb	0.0027851841588802

there is no presence of vulgarity among the bot accounts, while non-bot accounts may or may not include vulgar language.

Taking all this into account, a starting set of content features that may be selected for modeling may involve the following nine features: redund, urls, good_len, adj, tweets, vulgar, commnoun, art, emo_chat.

4.2 Limitations

A number of significant limitations must be noted.

First, the data set may not be a representative sample of the current state of affairs when it comes to bot versus non-bot activity in the Twitter medium.

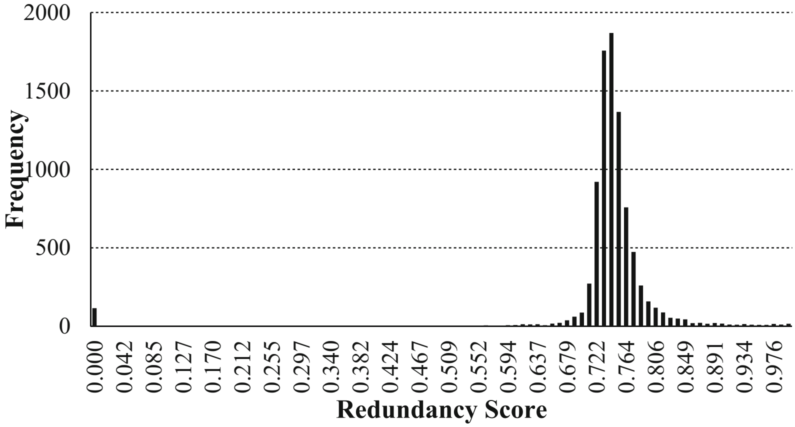

Fig. 1. Histogram of the distribution of redundancy score

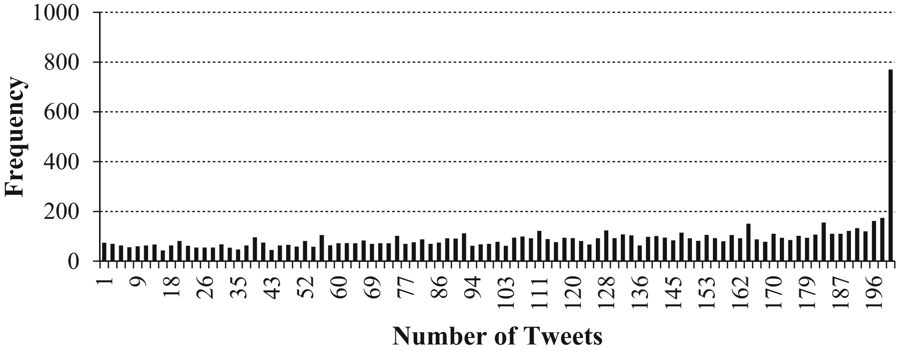

Fig. 2. Histogram of the distribution of number of tweets

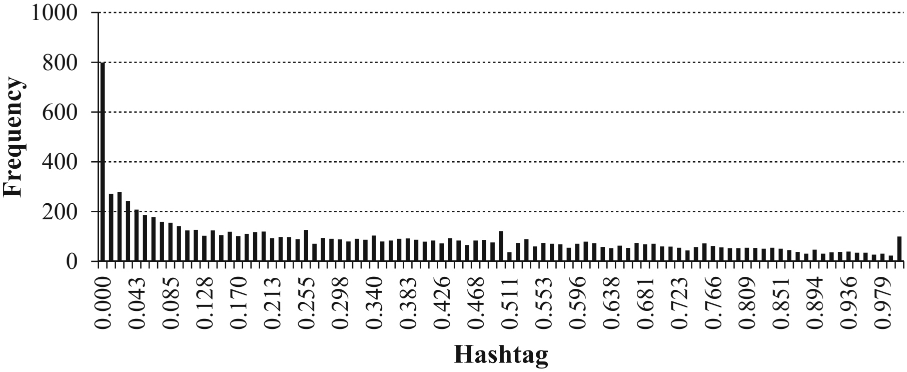

Fig. 3. Histogram of the distribution of hashtag

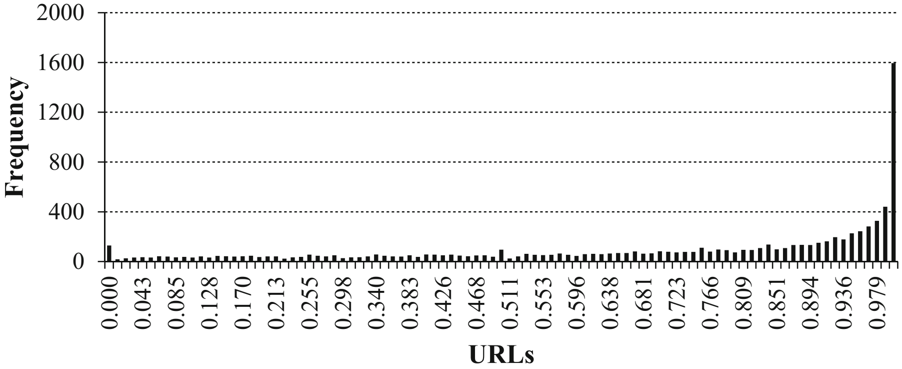

Fig. 4. Histogram of the distribution of URLs

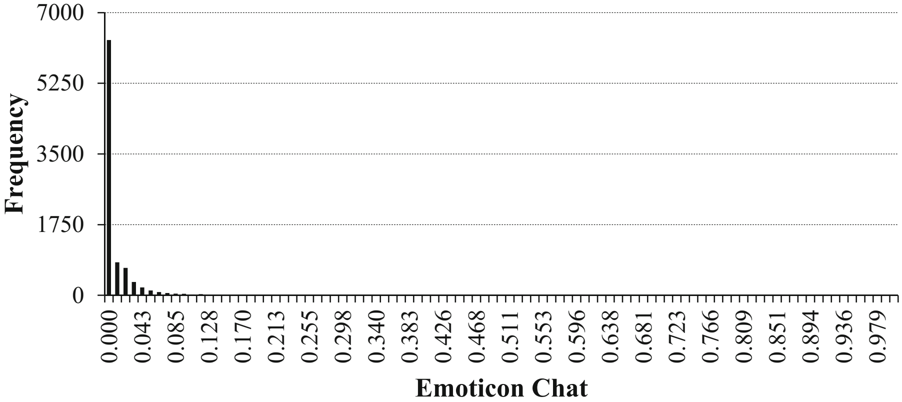

Fig. 5. Histogram of the distribution of emoticon_chat

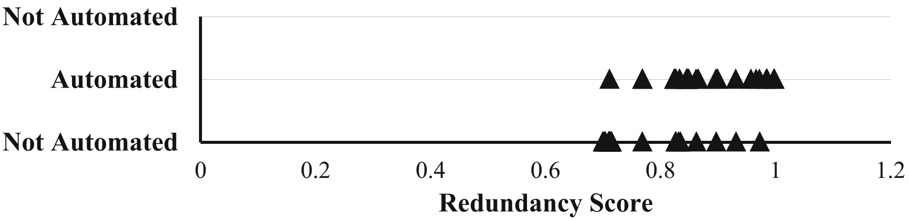

Fig. 6. Scatter plot of dependent variable against redundancy score

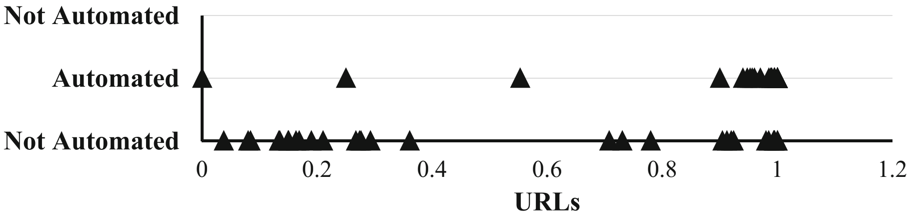

Fig. 7. Scatter plot of dependent variable against URLs score

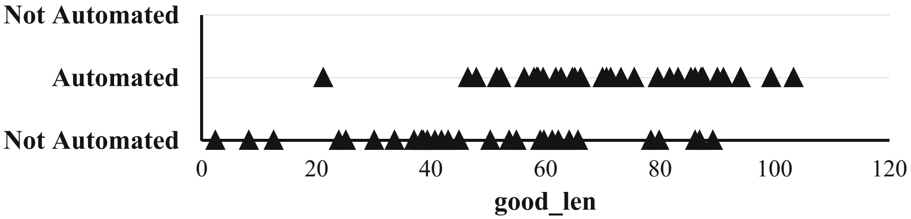

Fig. 8. Scatter plot of dependent variable against "good_len" score

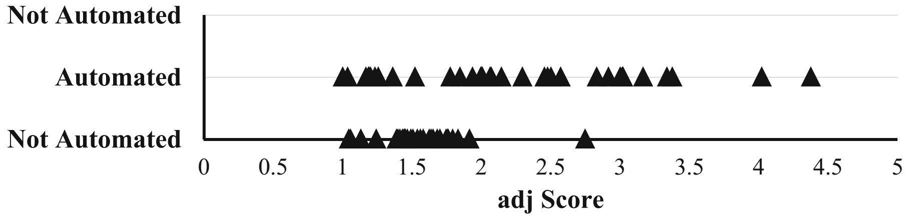

Fig. 9. Scatter plot of dependent variable against "adj" score

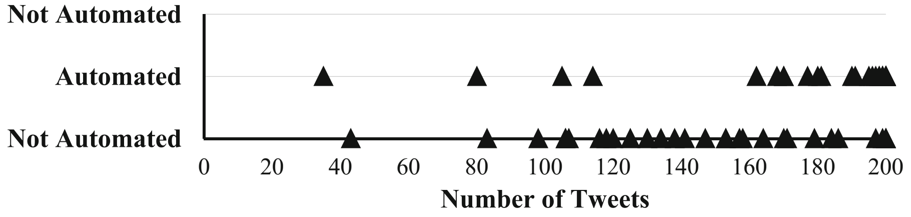

Fig. 10. Scatter plot of dependent variable against number of tweets

Second, the process of manually classifying a small set of accounts and reaching a consensus in roughly two-thirds of the cases may not be without errors.

Third, a larger sample set from the manual classification process may lead to different conclusions about content features and the type of modeling that may be expected to perform best.

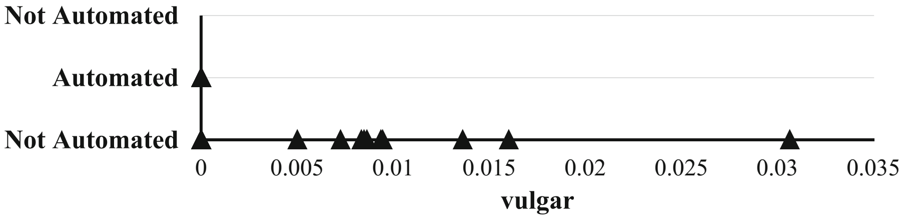

Fig. 11. Scatter plot of dependent variable against vulgarism score

Fourth, concentrating on content, which probably provides the most predictive power, may still ignore some critical external features, and thus may not produce an optimal perspective.

4.3 Further Investigations

Future work may attempt to consider a mix of external features and content features, calculated on a large set of known bot and non-bot accounts for better feature selection, description, and classification. This should enable a much more reliable subset of predictive or discriminating features, which in turn may lead to more reliable descriptive and predictive models.

5 Conclusion

This paper demonstrates one way by which content of social media activities may be processed in terms of mathematical "signatures" of different types of online behaviors that may be used for descriptive and predictive modeling of automated versus non-automated activities.

References

1. Alarifi, A., Alsaleh, M., Al-Salman, A.: Twitter turing test: identifying social machines. Inf. Sci. **372**, 332–346 (2016). doi:10.1016/j.ins.2016.08.036
2. Carapinha, F., et al.: Modeling of social media behaviors using only account metadata. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2016. LNCS (LNAI), vol. 9744, pp. 393–401. Springer, Cham (2016). doi:10.1007/978-3-319-39952-2_38
3. Chu, Z., Gianvecchio, S., Jajodia, S., Wang, H.: Detecting automation of Twitter accounts: are you a human, bot, or cyborg? IEEE Trans. Dependable Sec. Comput. **9**, 811–824 (2012)
4. Dickerson, J.P., Kagan, V., Subrahmanian, V.: Using sentiment to detect bots on Twitter: are humans more opinionated than bots? In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) (2014). doi:10.1109/asonam.2014.6921650

5. A framework for twitter bot analysis. In: Proceedings of the 25th International Conference Companion on World Wide Web - WWW 2016 Companion (2016). doi:10.1145/2872518.2889360
6. Main, W., Shekokhar, N.: Twitterati Identification System (2015). http://www.sciencedirect.com/science/article/pii/S1877050915003129. Accessed 29 Jan 2017
7. Hancock, M.: Automating the characterization of social media culture, social context, and mood. In: 2014 Science of Multi-Intelligence Conference (SOMI), Chantilly, VA (2014)
8. Hancock, M., Sessions, C., Lo, C., Rajwani, S., Kresses, E., Bleasdale, C., Strohschein, D.: Stability of a type of cross-cultural emotion modeling in social media. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) AC 2015. LNCS (LNAI), vol. 9183, pp. 410–417. Springer, Cham (2015). doi:10.1007/978-3-319-20816-9_39