

An Evaluation of Low-Quality Content Detection Strategies: Which Attributes Are Still Relevant, Which Are Not?

Júlio Resende[1], Vinicius H. S. Durelli[1], Igor Moraes[1], Nícollas Silva[2], Diego R. C. Dias[1(✉)], and Leonardo Rocha[1]

[1] Universidade Federal de São João del-Rei, São João del-Rei, Brazil
{julio.cmdr,durelli,ti.igor,diegodias,lcrocha}@ufsj.edu.br
[2] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
ncsilvaa@dcc.ufmg.br

Abstract. Online social networks have gone mainstream: millions of users have come to rely on the wide range of services provided by social networks. However, the ease use of social networks for communicating information also makes them particularly vulnerable to social spammers, i.e., ill-intentioned users whose main purpose is to degrade the information quality of social networks through the proliferation of different types of malicious data (e.g., social spam, malware downloads, and phishing) that are collectively called low-quality content. Since Twitter is also rife with low-quality content, several researchers have devised various low-quality detection strategies that inspect tweets for the existence of this kind of content. We carried out a brief literature survey of these low-quality detection strategies, examining which strategies are still applicable in the current scenario – taken into account that Twitter has undergone a lot of changes in the last few years. To gather some evidence of the usefulness of the attributes used by the low-quality detection strategies, we carried out a preliminary evaluation of these attributes.

Keywords: Low-quality content detection · Spam · Machine Learning

1 Introduction

Over time, what we have referred to as the Web has undergone several gradual changes. Specifically, the Web started as static HTML pages (i.e., read-only) and evolved to a more dynamic Web (i.e., read-write): Web 2.0 is the term used to refer to this ascendant technology whose emphasis lies in allowing users to create content, collaborate, and share information online. Owing to its interactive and dynamic nature, a great deal of sites that lend themselves well to the Web 2.0 model have been developed with an interactive community of users in mind. Quintessential examples of social networking services include Facebook and Twitter. These websites have become increasingly popular because they

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12249, pp. 572–585, 2020.
https://doi.org/10.1007/978-3-030-58799-4_42

provide users with community-building options such as online forums for like-minded people. These socially-enabled websites have been playing an active role in allowing individuals and communities to make more informed decisions concerning a wide range of topics. Despite the many benefits offered by these social networking services, online social networks also have issues. Due to the massive amount of information available in social networking services, these websites are particularly vulnerable to common approaches to tampering with the quality of the information. According to [8], anything that hinders users from having a meaningful browsing experience can be regarded as low-quality content. Thus, social spam, phishing, and malware, which are common approaches to degrading information, are examples of low-quality content that can negatively affect the overall user experience provided by social networking services.

The motivation for propagating low-quality content ranges from product advertising to socially engineered malware (i.e., malware requiring user interaction that ends up with users downloading a malicious file). Additionally, it is worth noting that low-quality content can be propagated either by users or social *bots*. In effect, bots are estimated to make up more than 50% of all Internet traffic [29]. In this context, given that bots are able to propagate a massive amount of low-quality content and the information in social networking services plays a pivotal role in shaping the decisions of users, bot traffic has the potential to severely user experience. Twitter stands out as one of the social networking services that is most vulnerable to spam: out of approximately 700 million *tweets*, roughly 10% is considered spam [23]. The reasons for Twitter being plagued by low-quality content are manyfold, but the main reason has to do with the fact that the service places too high of a limit on the number of tweets per day[1]

Due to the aforementioned reasons, the detection of low-quality content has been drawing a lot of attention from researchers and practitioners. However, in the literature is presented a limited interpretation of how to detect low-quality content. Typically, the most common interpretations emphasize the detection of spammer accounts or the detection of malicious links, which corroborates the generation of incomplete filters. False information (i.e. fake news), automatically generated content, flood, clickbait and advertisements are also data that do not contain any information and can also be published by legitimate accounts, not being treated by these strategies. Another problem is that many approaches have several historical resources to perform the detections, such as a user's old tweets or his progression in terms of the number of followers, some of which are no longer available due to restrictions imposed by the new Twitter policy.

In this sense, in hopes of providing an overview of the main approaches researchers and practitioners have been investigating to cope with this problem, we set out to give a survey of the scope of the literature. While conducting such a literature survey, we emphasize checking which strategies and resources (i.e., attributes) remain viable to address the problem of low-quality content

[1] As of this writing, the tweet limit is 2,400 per day. It is worth emphasizing that there is also a limit per half-hour period, so users are not able to tweet all 2,400 tweets at one time.

online detection. Additionally, we came up with a methodology that combines different low-content detection techniques. Essentially, our approach was devised as a means to evaluate the effectiveness of some prominent low-content detection classification models. Thus, our contributions are twofold:

1. We carried out a brief literature survey of low-quality content detection techniques. Our survey was conducted from the standpoint of a researcher investigating which techniques are still applicable in the current scenario – considering that social networking services have evolved over the last few years.
2. To gather evidence of the usefulness of some of the low-quality content detection techniques, we devised a methodology that combines them so that we can probe into their advantages and drawbacks in terms of the attributes that they employ.

The remainder of this work is organized as follows. In Sect. 2 we present the main proposed approaches for detecting low-quality content, emphasizing which of them remain viable to be applied in online scenarios. In Sect. 3, we describe the methology used to evaluated the strategies and characteristics currently being used for low-quality content online detection. In Sect. 4, we present the results and discussions inherent in this work. Finally, in Sect. 5, we summarize our conclusions.

2 Literature Survey

Several articles use the same terminology - spam - to refer to different concepts. The confusion occurs because this term is commonly used to refer to one of the first types of spam to be popularized on the internet, the "phishing spam".

As the term is ambiguous and "Low-quality Content" is a terminology that is not well established in the literature, we decided to follow a "quasi-gold standard" (i.e., a manual search of several articles in the subject area). The quasi-gold standard was used to define a starting point for a "snowballing" technique. A pre-filter was done through the abstract of several articles, and later we made a second one, reading the entire research.

In this way, we selected [8], our starting point for the literature survey. From it, the "backward snowballing" technique was used (i.e., all articles cited by him were analyzed). We focused our research on low-quality content, so the technique should be reapplied in articles dealing with this type of detection. For articles that deal with only one type of low-quality content, we limit it to just citing them. The diagram in Fig. 1 demonstrates the process.

Only one new article [25] was found that deals with low-quality content (although it used the term "spam"), so we repeated in this article the process of backward snowballing, of which no study was found that follows the methodology sought.

After the "backward snowballing", we used Google Scholar to explore articles that cited the relevant papers, technique known as "forward snowballing". However, the search did not return any new study relevant to our survey, and

we decided not to include them in this article because they did not present any new approach in relation to the citations selected by the previous technique.

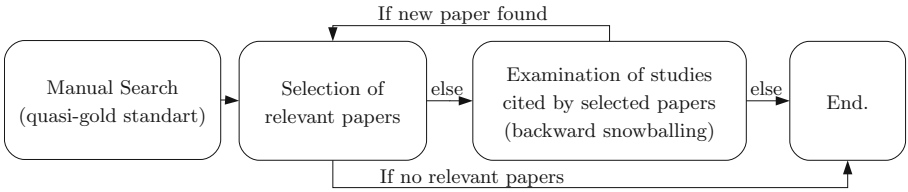


Fig. 1. Backward snowballing diagram

Low-quality content or spam on social networks is unwanted content produced by sources that express different behavior from that intended, whether that source is a legitimate user, a spammer, or a bot [25]. The purpose of the content includes sharing malicious links, fraudulent information, advertisements, chain letters, and others, as well as indexing of unrelated hashtags to increase visibility. However, we note that most published works do not share the same definition, using only a narrow view, which does not represent the entire term. The most modern approach in the literature is to consider only phishing spam, a strategy of spreading malware through URLs. The detection of this content was the object of study of [1,15], which proposed techniques to verify the occurrence of the registered blacklist (i.e., an extensive catalog of suspicious or malicious URLs) and techniques which learn about the characteristics of these blacklists, in order to apply the knowledge generated to classify new URLs. A negative point of these strategies is that spammers may shorten the link or even generate a new one to circumvent this detection. [18] has improved this methodology by measuring the similarity of words in tweets, in a similar way as [3] detects evidence of crimes in forums. Thus, if a tweet has an unknown URL or text content similar to other tweets with suspicious URLs, it is detected as phishing. Twitter itself uses this methodology, checking links with Google Safe Browsing. However, shortened malicious links are still not detected as spam. The tweets barred are full links, those identical to those already published in a short period of time, and those in which the sender and receiver have never interacted with each other.

Considering as a definition of spam, only messages containing phishing is a limited approach. Seeking to broaden the detection spectrum, [7] based its definition on reports made by other users, a better approach if the reports would be publicly available. Another approach, explored by [11,20,22], was the creation of a database to detect tweets, labeling them as spam if Twitter suspended the sender account in a later validation request. However, relying on Twitter's suspension policy is not the best way, since posting spam is not the only reason for an account to be suspended. Besides, spammers can post regular tweets to avoid being detected, while legitimate users can post spam content, even if

unconsciously. Other works that also focus on the detection of spammer accounts, such as [2,16,17], use account attributes extracted through Twitter's own API's to generate classifying models with machine learning. However, the focus on these attributes can bias the search, since external tools can easily manufacture them, such as obtaining followers or scheduled publications.

The generation of temporal behavior models, explored by [6,10,12,13,21,26–28], is an alternative developed to detect inorganic variations of these attributes values. Considering historical user data, the models detect anomalies and then flag the account as a spammer. The approach is very efficient in detecting spammers but does not consider those legitimate users also to post spam. Another hindrance is that the use of time resources infers that spammers are only detected after sending several spam messages, as this type of detection requires a minimum amount of time for the analyzed accounts to generate "evidence" that proves illegal activity. There are graph-based detection methodologies, which seek to trace the entire communicative path taken by users and can also detect communities [4]. An interesting work is presented by [5], in which, by calculating the centrality network, the largest spammers can be detected. Similar methods are used by [9,19,24], in which the extracted attributes determine the number of interactions made per user, as well as the distance between the sender and the receiver in targeted tweets. That significantly improves accuracy at the cost of response time. In the same line of temporal behavior, we can cite the work presented by [10]. The proposed methodology is the clustering of tweets that use the same URL (presenting an overlap with the group of papers described in the first paragraph of this section), being, therefore, a phishing detection. Later, using blacklists, the group that presents a malicious URL is considered a spam campaign, that is, mass dissemination of the same malicious content. In other words, all these strategies take some time to identify inappropriate behavior. Due to the damage caused by these contents, it is of utmost importance to consider methodologies that perform the detection in real-time.

Assuming as an ideal methodology: (1) the detection of the various types of low-quality tweets; (2) being this detection in real-time; and (3) using publicly available attributes – we found only two articles that meet all these requirements [8,25]. [25] proposed a strategy to detect spam in its broadest sense. However, the dataset used for evaluation was composed only of tweets that presented at least one URL. It is, therefore, not possible to measure its ability to detect tweets that do not have this attribute. On the other hand, [8] has labeled a database according to users' perspectives on what is low-quality content. That is an exciting new starting point, given the abstract definition of what is "unwanted or irrelevant" content. Comparing this work with others that adopt a narrow definition of what is low-quality content is not statistically valid, because the works have different focuses. However, [8] presented a comparison of results with those presented on [25], and the results of [8] proved to be more efficient. A critical point of [8], regarding the attributes used to build the classification model, was the use of some attributes that should not be considered in real-time detection, such as the number of retweets and favorites received by

the tweet. As detection should be applied when publishing a tweet, this value will always be 0. The same work also proposes a methodology that makes use of attributes called indirect, which depend on an analysis of all the posts made by the user. Although this methodology is capable of significantly increasing the performance of the classifiers, the excessive use of the user's information has not been seen with good eyes in recent years, a fact that has pressured Twitter to reduce the limits for obtaining information.

Given the analysis prepared, the next section presents a methodology for evaluating the attributes used by the works that fit the context defined by this research [8,25]. The objective was to analyze the relevance of each attribute and the feasibility of carrying out a broad detection of low-quality contents.

3 Evaluation Methodology

As mentioned, we set out to investigate the importance of techniques currently being used for low-quality content detection. We propose an evaluation methodology to be applied to the main strategies proposed in the literature. To this end, we tried to use a manually labeled Twitter dataset comprised of 100,000 tweets generated by 92,720 different users, made available by Chen et al. [8]. According to Chen et al., the tweets were manually verified so as to provide a more sound training set and allow for the verification of the results. Nevertheless, recently, Twitter has introduced a number of features that prevent researchers from directly accessing tweet datasets. To work around this limitation, we took advantage of the fact that all tweets' IDs in the aforementioned dataset were made available by Chen et al., thus we collected all tweets' information based on their IDs. These were the only data collected for our analysis. During the creation of our dataset, we realized that some tweets were no longer available, so we ended up with a dataset containing 43,857 tweets, of which 3,967 were classified as low-quality. It is worth mentioning that although our new dataset is smaller and different from the original, the ratio between low-quality and normal messages was maintained.

During the creation of the dataset, apart from the tweets' textual content, we also extracted information concerning other attributes. As mentioned, attributes related to previous tweets and followers can be used as useful pieces of information, improving the models yielded by the classifiers. We decided not to include information regarding retweets, previous tweets, and the number of times a given tweet was liked/favorited because Twitter has introduced features that make the extraction of such information somewhat unwieldy. Besides, the use of this information in low-quality content online detection scenarios has limited utility. Therefore, the attributes we took into account are shown in Table 1. Our selection of attributes is based on the selection proposed by other researchers: attributes 3, 4, 6, 7, 8, 9, 15, 16, 17, 18, and 35 were used in the approach proposed by [8], attributes 12, 13, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 32, and 34 were selected and used in the approach devised by [25], and 1, 2, 5, 10, 11, 29, 31 and 33 are attributes employed in both approaches. It is worth noting that

Table 1. Set of attributes we took into account during our study. The attributes 3, 4, 6, 7, 8, 9, 15, 16, 17, 18, and 35 were used in the approach proposed by [8], attributes 12, 13, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 32, and 34 were selected and used in the approach devised by [25], and 1, 2, 5, 10, 11, 29, 31 and 33 are attributes employed in both approaches.

Id	Description	Id	Description
1	# of followers	19	# of words in the tweet
2	# of followings	20	# of characters in the tweet
3	# of lists the user create	21	# of digits in the tweet
4	# of tweets the user favorite	22	# of white spaces in the tweet
5	# of tweets the user posted	23	# of capitalization words in the tweet
6	If the URL field of profile is null	24	$\frac{Attribute23}{Attribute19}$
7	Length of the description field of profile	25	Length of longest word in the tweet
8	If the user is using a default profile	26	Average word length in the tweet
9	If the user is using a default avatar	27	# of exclamation marks in the tweet
10	Age of the user account, in months	28	# of question marks in the tweet
11	$\frac{Atribute1}{Atribute2}$	29	# of urls in the tweet
12	$\frac{Atribute1}{Atribute1+Atribute2}$	30	$\frac{Attribute29}{Attribute19}$
13	Average followings per month	31	# of hashtags in the tweet
14	Average tweets per week	32	$\frac{Attribute31}{Attribute19}$
15	If the location field of profile is null	33	# of mentions in the tweet
16	Tweeting tools	34	$\frac{Attribute33}{Attribute19}$
17	Regular, replies mentions or retweets	35	# of symbols in the tweet
18	Tweet contains sensitive content		

many other approaches to spam detection (i.e., low-quality content detection) consider a subset of the set of attributes we took into account in our study.

Many of the attributes listed in Table 1 can not be obtained through Twitter's API. Rather, some attributes have to be computed from other information. Attribute 17, for instance, can be inferred from boolean attributes that indicate whether a tweet is a "regular" tweet, a retweet, a (public) reply, or a mention (a special case of reply in which users mention other users explicitly, i.e., `@username`). Attributes 19, 20, 21, and 22 can be extracted by applying regular expressions to the tweet's textual information.

Special attention was given to attribute 16, which indicates the medium used to post the tweet (e.g., an iPhone app). During data collection we found that such attributes can have more than one hundred different values. None of the previous work on low-content detection mention the approach they used to choose and interpret these values, so we had to devise our own approach to doing such. We calculate the frequency of these values both in instances classified as low-content as well as "regular" tweets. Afterwards, we calculated the module

of the difference of each frequency for both types of instances. The calculated values were then ordered in descending order. In order to choose the values to be considerated in our analysis, we applied an interpolation function to the difference of the frequencies. We came up with a cut-off value based on the ΔY of interpolation function, considering the point at which the function decreases. As a result, we selected the following values: Twitter for iPhone, Twittascope, and Twitter for Android. Values that did not fall into these categories were classified as "Others". Figure 2 shows the frequency probability of each category for low-quality content and regular tweets.

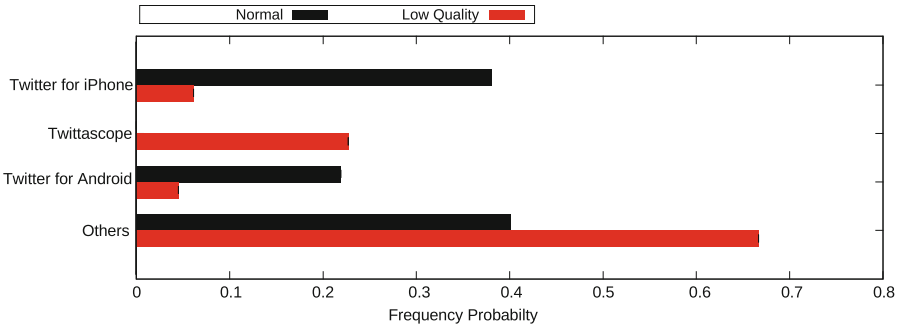


Fig. 2. Frequency probabilty of possible values for attribute 16. For instance, for normal tweets, 38% of them are posted from a Iphone, 22% from Androide devices and the remaining of the tweets posted by other types of devices. Moreover, more than 60% of low-quality messages are posted by other types of device.

As mentioned, we set out to examine which attributes are more relevant, i.e., attributes that contribute the most to the prediction of low-quality content. To this end, we carried out an analysis based on feature selection algorithms: *Chi2* [14] and *Information Gain*. At a basic level, both algorithms assign weights to each attribute, from most relevant to model creation to least relevant. In this context, relevance is measured in terms of how much a given attribute contributes to the resulting model. Therefore, we needed to adopt an additional strategy to select relevant attributes based on their respective weights. Specifically, we also needed to decide how many attributes should be chosen, and which combination most favors model construction. Over the course of our analysis, we investigated all possible sizes of subset of attributes (i.e. the most important attribute, the two most important attributes, the three most important attributes, etc.). For each set of attributes, we construct a classification model. For the classification model, we consider the Random Forest algorithm, since it has been widely used in the literature. To validate our results, we used *F-measure* as the metric and a 10 fold cross-validation strategy.

To define the F-Measure metric, we need to understand two main concepts:

– Precision: number of items classified as positive is positive;

– Recall: number of relevant items selected.

The F-Measure (**F1**) is the harmonic mean between precision and recall:

$$\mathrm{F1} = 2 * \frac{\mathrm{precision} * \mathrm{recal}}{\mathrm{precision} + \mathrm{recall}} \tag{1}$$

In the end, a comparison of the most efficient model resulting from tests that take into account the methods of ranking attributes (described in the last paragraph) with three other data sets was performed. Initially, we fed all 35 attributes shown in Table 1 into two algorithms: RF and Support Vector Machine (SVM). We also tried a combination of attributes that emphasize all user-related information (attributes 1 to 14). The third set is centered around content-related attributes (attributes 15 to 35). We used *F-measure* as metric and a 10-fold cross-validation strategy to probe into the effectiveness of the resulting models.

4 Analysis and Results

As mentioned in the previous section, after applying Chi2 and Information Gain, we obtained two lists of attributes ordered by weight assigned by each feature selection algorithm. The results of applying both algorithms are shown in Fig. 3. To better present the results, we normalized the weights to correspond to a scale from 1 to 100. Our results would seem to suggest that attribute 16, which represents the medium used to post the tweet, is the attribute that best contributes to create a classification model. In way, this was expected since in our dataset all content posted through Twittascope was labeled as low-quality content (as shown in Fig. 2) [8].

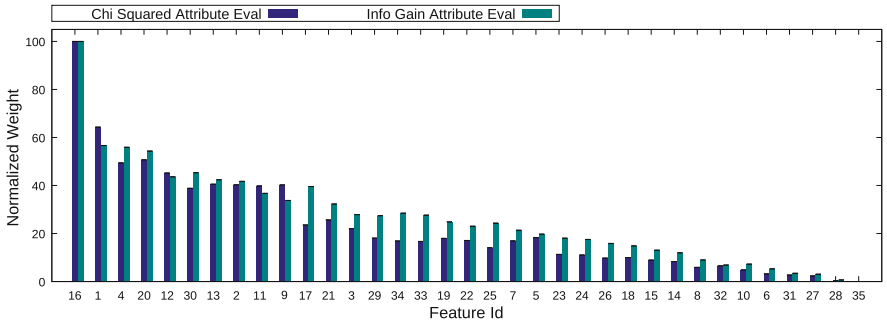


Fig. 3. Score of each attribute generated by the feature selection algorithms *Chi2* e *Information Gain*. While attribute 16 is the one that best contributes for creating a classification model, the attributes 28 and 35 are the worst.

Based on these two lists od attributes, we perform the second step of our methodology to evaluate the subset of attributes that can be used to create the

best low-quality content detection model. As previously mentioned, we create models adopting the Random Forest algorithm and considering different sizes of attributes (i.e. the most important attribute, the two most important attributes, the three most important attributes, etc.). Figure 4 shows how much the predictive power of the models increases in response to the number of attributes taken into consideration during model generation. Upon analyzing Fig. 4, it can be seen that the predictive power of the resulting models peaked (i.e., achieved the highest F-measure score) when 33 attributes were fed into both algorithms: in effect, the most effective predictive models were generated when the subset of attributes does not include attributes 28 and 35. However, it interesting to note that with the 13 most relevant attributes we achieve a quality very close to the highest F-measure considering 33 attributes. The 13 most relevant attributes are 16, 1, 4, 20, 12, 30, 13, 2, 11, 9, 17, 21 and 3. As future work, we intend to deep investigate these attributes.

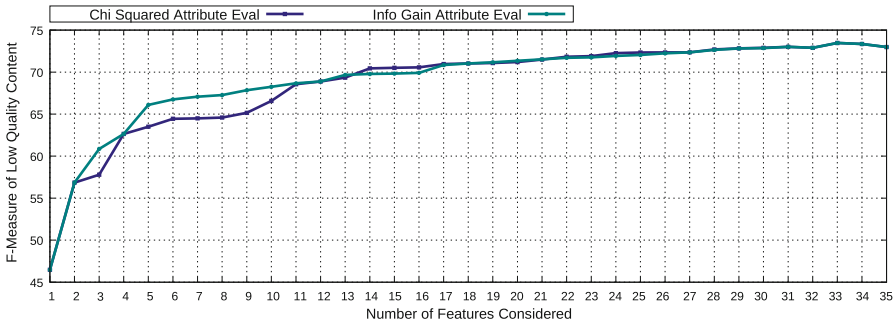


Fig. 4. F-Measure da classe conteúdos de baixa qualidade para variadas quantidades de atributos segundo os algoritmos *Chi2* e *Information Gain*

Table 2 presents more comprehensive information on how each algorithm performed when trained with each of the aforementioned attribute sets. We evaluated the resulting models in terms of two proxy metrics for their prediction power: true positive rate (TPR) and F1. Moreover, Table 2 also lists the accuracy and macro-F1 of the generated predictive models. Accuracy measures the global effectiveness regarding all decisions made by the classifier (that is, the inverse of error rate). Macro-F1 on the other hand, measures the classification effectiveness regarding each class independently. It corresponds the mean of the F-Measure values obtained for each possible class in the dataset.

Random Forest yielded better models for all subsets of attributes when compared to SVM (Table 2). A more straightforward interpretation of the results indicates that all subsets achieved over 90% accuracy for all algorithms. The best performing combination of attributes comprises 33 attributes that represent user and tweet information. The second best performing attribute set includes only attributes related to tweets. Our results would seem to suggest that using

Table 2. Classification Effectiveness considering different set of attibutes. Random Forest yielded better models for all subsets of attributes when compared to SVM. The best results are achieved considering the 33 most relevant attibutes using the Random Forest algorithm. Moreover, the set of attributes containing only user information does not yield good low-quality content predictive models.

Set of attributes	Classifier	Accuracy	TPR	F1	MacroF1
33 most relevant attributes	SVM	93.66	43.31	55.28	75.93
	Random Forest	**96.15**	60.43	**73.81**	**85.87**
All Attributes	SVM	93.6	43.31	55.28	75.93
	Random Forest	96.0	59.57	73.09	85.47
User's Attributes (1 - 14)	SVM	91.48	14.84	22.68	59.01
	Random Forest	93.50	36.45	50.36	73.44
Tweet's Attributes (15 - 35)	SVM	94.95	45.27	61.87	79.58
	Random Forest	95.77	**60.73**	72.19	84.95

a set of attributes containing only user information does not yield good low-quality content predictive models. This is key confirmatory finding, we argue that researchers and practitioners can refrain from including user information in their low-quality content prediction approaches since this type of information does not seem to contribute much to the creation of effective models.

5 Concluding Remarks

Owing to the popularity of social networking services, low-content quality has become an active research area and has been drawing the attention of researchers and practitioners alike. We carried out a literature survey in hopes of giving an overview of the most prominent research efforts to advance understanding in the low-quality content area. We believe our survey can be useful for researchers and practitioners looking to get a better understanding of what has already been investigated in the area. As for practitioners, our survey can be seen as an initial foray into promising approaches to low-quality content detection.

We also focused on examining the effectiveness of the attributes researchers have been employing to support low-quality content approaches. In the context of our research, effectiveness can be seen in terms of how much attributes are able to contribute to the generation of predictive models, that is, how much attributes contribute to the whole low-quality content detection strategy in question. For clarity purposes, we framed our discussion and evaluation of such attributes in terms of two categories: *(i)* user-related attributes and *(ii)* content-based attributes. A manually labeled Twitter dataset, which was employed in a previous study [8], was used in our evaluation. According to the results of such evaluation, models generated using only content-based attributes are as good as the models using the 33 best attributes. This could be evidenced by the TPR

metric, in which both models had a score of approximately 60.5% using the Random Forest classifier. In contrast, models based on only user-related information had a score of 36.45% for the same metric and classifier. As mentioned, this is a fundamental confirmatory finding because it provides evidence that researchers and practitioners can refrain from including user information in their low-quality content prediction strategies without having to compromise on prediction ability.

As future work, we plan to come up with new attributes that can be used to yield better predictive models. Also, we intend to evaluate the performance of the resulting models using datasets from different social networking services and not only Twitter. We also plan to investigate different strategies for attribute selection. For instance, we believe that the Wrapper method might be a promising approach to selecting attributes that can be fed into some algorithms that have been widely explored in recent years, e.g., Convolutional Neural network (CNN).

Acknowledgments. This work was partially funded by the Brazilian National Institute of Science and Technology for the Web - INWeb, MASWeb, CAPES, CNPq, Finep, Fapesp and Fapemig.

References

1. Aggarwal, A., Rajadesingan, A., Kumaraguru, P.: PhishAri: automatic realtime phishing detection on twitter. In: 2012 eCrime Researchers Summit. IEEE, October 2012. https://doi.org/10.1109/ecrime.2012.6489521
2. Almaatouq, A., et al.: Twitter: who gets caught? Observed trends in social microblogging spam. In: Proceedings of the 2014 ACM conference on Web science - WebSci. ACM Press (2014). https://doi.org/10.1145/2615569.2615688
3. Azeta, A.A., Omoregbe, N.A., Ayo, C.K., Raymond, A., Oroge, A., Misra, S.: An anti-cultism social education media system. In: 2014 Global Summit on Computer & Information Technology (GSCIT). IEEE, June 2014. https://doi.org/10.1109/gscit.2014.6970097
4. Behera, R., Rath, S., Misra, S., Damaševičius, R., Maskeliūnas, R.: Large scale community detection using a small world model. Appl. Sci. **7**(11), 1173 (2017). https://doi.org/10.3390/app7111173
5. Behera, R.K., Rath, S.K., Misra, S., Damaševičius, R., Maskeliūnas, R.: Distributed centrality analysis of social network data using MapReduce. Algorithms **12**(8), 161 (2019). https://doi.org/10.3390/a12080161
6. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on Twitter. In: Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)
7. Bosma, M., Meij, E., Weerkamp, W.: A framework for unsupervised spam detection in social networking sites. In: Baeza-Yates, R., et al. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 364–375. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28997-2_31
8. Chen, W., Yeo, C.K., Lau, C.T., Lee, B.S.: A study on real-time low-quality content detection on Twitter from the users' perspective. PLOS One **12**(8), 1–22 (2017). https://doi.org/10.1371/journal.pone.0182487

9. Fakhraei, S., Foulds, J., Shashanka, M., Getoor, L.: Collective spammer detection in evolving multi-relational social networks. In: Proceedings of the 21th SIGKDD. ACM Press (2015). https://doi.org/10.1145/2783258.2788606
10. Gao, H., Chen, Y., Lee, K., Palsetia, D., Choudhary, A.: Poster. In: Proceedings of the 18th ACM conference on Computer and communications security. ACM Press (2011). https://doi.org/10.1145/2046707.2093489
11. Hu, X., Tang, J., Gao, H., Liu, H.: Social spammer detection with sentiment information. In: 2014 IEEE International Conference on Data Mining. IEEE, December 2014. https://doi.org/10.1109/icdm.2014.141
12. Jin, X., Lin, C.X., Luo, J., Han, J.: Socialspamguard: a data mining-based spam detection system for social media networks. In: Proceedings of the International Conference on Very Large Data Bases (2011)
13. Lee, K., Eoff, B.D., Caverlee, J.: Seven months with the devils: a long-term study of content polluters on Twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
14. Liu, H., Setiono, R.: Chi2: feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society Press (1995). https://doi.org/10.1109/tai.1995.479783
15. Martinez-Romo, J., Araujo, L.: Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Syst. Appl. **40**(8), 2992–3000 (2013). https://doi.org/10.1016/j.eswa.2012.12.015
16. McCord, M., Chuah, M.: Spam detection on Twitter using traditional classifiers. In: Calero, J.M.A., Yang, L.T., Mármol, F.G., García Villalba, L.J., Li, A.X., Wang, Y. (eds.) ATC 2011. LNCS, vol. 6906, pp. 175–186. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23496-5_13
17. Miller, Z., Dickinson, B., Deitrick, W., Hu, W., Wang, A.H.: Twitter spammer detection using data stream clustering. Inf. Sci. **260**, 64–73 (2014). https://doi.org/10.1016/j.ins.2013.11.016
18. Santos, I., et al.: Twitter content-based spam filtering. In: Herrero, Á., et al. (eds.) International Joint Conference SOCO'13-CISIS'13-ICEUTE'13. AISC, vol. 239, pp. 449–458. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-01854-6_46
19. Song, J., Lee, S., Kim, J.: Spam filtering in Twitter using sender-receiver relationship. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 301–317. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23644-0_16
20. Sridharan, V., Shankar, V., Gupta, M.: Twitter games. In: Proceedings of the 28th ACSAC. ACM Press (2012). https://doi.org/10.1145/2420950.2421007
21. Tan, E., Guo, L., Chen, S., Zhang, X., Zhao, Y.: Spammer behavior analysis and detection in user generated content on social networks. In: 2012 IEEE 32nd International Conference on Distributed Computing Systems. IEEE, June 2012. https://doi.org/10.1109/icdcs.2012.40
22. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference. ACM Press (2011). https://doi.org/10.1145/2068816.2068840
23. Ungerleider, N.: Almost 10% of Twitter is spam (2015). https://www.fastcompany.com/3044485/almost-10-of-twitter-is-spam. Accessed 02 July 2019
24. Wang, A.H.: Don't follow me: spam detection in twitter. In: 2010 International Conference on Security and Cryptography (SECRYPT), pp. 1–10, July 2010

25. Wang, B., Zubiaga, A., Liakata, M., Procter, R.: Making the most of tweet-inherent features for social spam detection on Twitter. arXiv preprint arXiv:1503.07405 (2015)
26. Yang, C., Harkreader, R., Gu, G.: Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans. Inf. Forensics Secur. **8**(8), 1280–1293 (2013). https://doi.org/10.1109/tifs.2013.2267732
27. Yang, C., Harkreader, R.C., Gu, G.: Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) RAID 2011. LNCS, vol. 6961, pp. 318–337. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23644-0_17
28. Zheng, X., Zhang, X., Yu, Y., Kechadi, T., Rong, C.: ELM-based spammer detection in social networks. J. Supercomput. **72**(8), 2991–3005 (2015). https://doi.org/10.1007/s11227-015-1437-5
29. Łuksza, K.: Bot traffic is bigger than human. make sure it doesn't affect you! (2018). https://voluum.com/blog/bot-traffic-bigger-than-human-make-sure-they-dont-affect-you/