TubeSpam: Comment Spam Filtering on YouTube

Conference Paper · December 2015

DOI: 10.1109/ICMLA.2015.37

3 authors, including:

Johannes V. Lochter
Universidade Federal de São Carlos
7 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Tiago A. Almeida
Universidade Federal de São Carlos
72 PUBLICATIONS **1,243** CITATIONS
SEE PROFILE

TubeSpam: Comment Spam Filtering on YouTube

Túlio C. Alberto, Johannes V. Lochter, Tiago A. Almeida
Department of Computer Science
Federal University of São Carlos – UFSCar
18052-780, Sorocaba, São Paulo, Brazil
{tuliocasagrande,jlochter}@acm.org, talmeida@ufscar.br

***Abstract*—The profitability promoted by Google in its brand new video distribution platform YouTube has attracted an increasing number of users. However, such success has also attracted malicious users, which aim to self-promote their videos or disseminate viruses and malwares. Since YouTube offers limited tools for comment moderation, the spam volume is shockingly increasing which lead owners of famous channels to disable the comments section in their videos. Automatic comment spam filtering on YouTube is a challenge even for established classification methods, since the messages are very short and often rife with slangs, symbols and abbreviations. In this work, we have evaluated several top-performance classification techniques for such purpose. The statistical analysis of results indicate that, with 99.9% of confidence level, decision trees, logistic regression, Bernoulli Naïve Bayes, random forests, linear and Gaussian SVMs are statistically equivalent. Based on this, we have also offered the TubeSpam – an accurate online system to filter comments posted on YouTube.**

I. Introduction

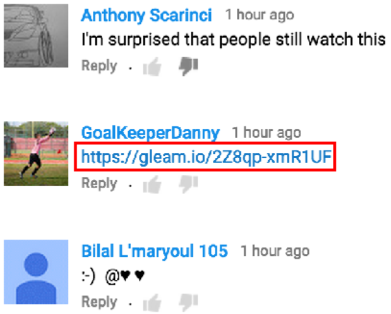
The popularization of broadband around the world has boosted the amount of Internet users. With faster connections, video host and sharing services became popular among users. According to a press release of Sandvine[1], a company focused on standards-compliant network policy control, around 55% of downstream traffic from United States is due to video platforms like Netflix and YouTube.

The availability of resources through Internet and the broadband connections allowed the appearance of sophisticated new platforms. In this way, YouTube is a famous video content publication platform with social network features, such as support for posting text comments to provide interaction between producer (channel owner) and viewers.

The success of YouTube can be expressed through recent statistics reported by Google[2]: the platform has more than 1 billion users, 300 hours of video are uploaded every minute and it generates billions of views every day. Around 60% of a creator's views come from outside their home country and half of YouTube views are on mobile devices.

Recently, YouTube has adopted a monetization system to reward producers, stimulating them to make high quality original content and increasing the amount of visualizations. After the deployment of this system, the platform was flooded by undesired content, usually of low quality information known as *spam*.

Among different kind of undesired content, YouTube is facing problems to manage the huge volume of undesired text comments posted by users that aim to self-promote their videos, or to disseminate malicious links to steal private data. Figure 1 presents an example of comment spam posted in one of the most viewed videos of YouTube.



Anthony Scarinci 1 hour ago
I'm surprised that people still watch this
Reply ·

GoalKeeperDanny 1 hour ago
https://gleam.io/2Z8qp-xmR1UF
Reply ·

Bilal L'maryoul 105 1 hour ago
:-) @♥♥
Reply ·

Figure 1. Example of a comment spam posted in YouTube.

The spam found on YouTube is directly related to the attractive profit offered by the monetization system. According to a press release by Google, more than a million advertisers are using Google ad platforms, the mobile revenue on YouTube is up 100% year over year and the number of hours people are watching on YouTube each month is up 50% year over year. At the same time, according Nexgate, a computer security company, just in the first half of 2013, the volume of social spam increased 355%[3]. For each spam found on any social network, other 200 spams are found on Facebook and YouTube.

The problem became so critical that it motivated users to create a petition in 2012, in which they ask YouTube to provide tools to deal with undesired content[4]. In 2013, the YouTube official blog reported efforts to deal with undesired comments through recognition of malicious links, ASCII art detection and display changes to long comments[5]. However, many users are still not satisfied with such solutions. In fact, in 2014, the user "PewDiePie", owner of the most subscribed channel on YouTube (nearly 40 million subscribers), disabled comments on his videos, claiming most of comments are mainly spam and there is no tool to deal with them[6].

[1] *Sandvine's Global Internet Phenomena Report (2014)*. Available at https://goo.gl/QHE4AY, accessed in July 23, 2015.

[2] *YouTube – Statistics (2015)*. Available at https://goo.gl/ozUXMB, accessed in July 23, 2015.

[3] *State of Social Media Spam (2013)*. Available at http://goo.gl/bnkUhh, accessed in July 23, 2015.

[4] *Give YouTube Users Tools to Combat Video Spam (2012)*. Available at https://goo.gl/dWUXgC, accessed in July 23, 2015.

[5] *An update on YouTube comments (2013)*. Available at http://goo.gl/GDZuWw, accessed in July 23, 2015.

[6] *The Guardian – PewDiePie switches off YouTube comments: 'It's mainly spam' (2014)*. Available at http://goo.gl/AixgI1, accessed in July 23, 2015.

The problem caused by social spam began to be seriously discussed from 2010, but an earlier work is dated from 2005 [1]. However, undesired comments on YouTube still harm the platform's community, evidencing such problem requires attention and research.

Established techniques for automatic spam filtering have their performance degraded when dealing with YouTube's comments. It is mainly due to the fact that such messages are usually very short and rife with idioms, slangs, symbols, emoticons and abbreviations which make even tokenization a challenging task.

Given this scenario, this paper presents a comprehensive performance evaluation of several well-known machine learning techniques that can be applied to automatically filter such undesired messages. Our main goal is to find promising methods and settings that can be used in an online tool developed to detect undesired text comments posted on YouTube, besides to offer new public datasets and good baselines for future comparisons.

The remainder of this paper is organized as follows: in Section II, we briefly describe the related work available in the literature. Section III offers the datasets description, methods and main settings used in the experiments. In Section IV, we present the achieved results. Section V presents the proposed TubeSpam, an online tool to automatically filter comment spam on YouTube. Finally, Section VI describes the main conclusion and offers guidelines for future work.

II. Related works

Spam is usually related to undesired content with low quality information. They are commonly found as images, texts or videos, hindering visualization of interesting content. There are many researches related to spam in literature, such as *web spam* [2], *blog spam* [3], [4], *e-mail spam* [5], [6] and *SMS spam* [7], [8] filtering. In social networks, undesired messages are known as *social spam*.

Blog comment spam is the most similar scenario to that investigated in this paper. However, the most-known strategy to detect a blog spam comment usually is to find the best representation of language model in post publication, using that representation to filter less related comments to its original subject [1], [9]. Such strategy can not be applied on YouTube, since comments are related to a video content with small or no textual description, therefore language models can not be properly mapped from original publication.

YouTube also faces malicious users that publish low quality content videos, which it is known as *video spam*. There are some studies in literature to find efficient ways to handle this activity through classification methods and feature extraction from metadata, such as title, description and popularity numbers [10], [11].

Another common alternative is automatic blocking *spammers* – users that disseminate spam [12], [13]. However, unlike spam disseminated in other social networks and email [14], [15], the spam posted on YouTube is not usually created by bots, but posted by real users aiming self-promotion on popular videos. Therefore, such messages are more difficult to identify due its similarity to legitimate messages.

Automatic spam filtering is useful in other tasks as well. Severyn *et al.* [16] reported significant improvement of performance in opinion detection task, when spam samples were removed before training a classifier.

As noted by Bratko *et al.* [17], spam filtering task slightly differs from similar text categorization problems. They claim undesired messages have chronological order and their characteristics may change according to that. It also explains that cross-validation is not recommended, because earlier samples should be used to train the methods, while newer ones should be used to test them. Furthermore, in spam filtering, errors associated with each class should be considered differently, because a blocked legitimate message is worst than an unblocked spam.

III. Methodology

To give credibility to the found results and in order to make the experiments reproducible, we present in this section the settings used for each classification method, as well as general information about datasets and experimental methodology.

A. Datasets

We have collected and created five databases composed by real, public and non-encoded data directly extracted from YouTube through its API[7], in the first semester of 2015. We have selected five of the ten most viewed YouTube videos during the collection period.

Each sample represents a text comment posted in the comments section of each selected video. No preprocessing technique was performed. Subsequently, each sample was manually labeled as spam or legitimate (*ham*), using a collaborative tagging tool developed for this purpose, called *Labeling*[8]. The samples have associated a metadata information, such as the author's name and publication date, which have been preserved.

Table I presents the datasets collected and used in the experiments reported in this paper, along with the YouTube video *ID*, the amount of samples in each class and the total number of samples.

Table I. Datasets collected and used in the experiments.

Datasets	YouTube ID	# *Spam*	# *Ham*	Total
Psy	9bZkp7q19f0	175	175	350
KatyPerry	CevxZvSJLk8	175	175	350
LMFAO	KQ6zr6kCPj8	236	202	438
Eminem	uelHwf8o7_U	245	203	448
Shakira	pRpeEdMmmQ0	174	196	370

All datasets used in this work are publicly available at http://dcomp.sor.ufscar.br/talmeida/youtubespamcollection/.

B. Experimental methodology

For the experiments, we have first processed the datasets, in which, only the texts of comments were used, discarding

[7] *YouTube Data API*. Available at https://developers.google.com/youtube/v3, accessed in 11/06/2015.

[8] *Labeling*. Available at http://lasid.sor.ufscar.br/ml-tools/.

associated metadata. We employed the *bag-of-words* model and frequency representation, as described below.

Considering each message m comprising a set of terms (*tokens*) $m = t_1, ..., t_n$, wherein each term t_k corresponds to a word with two or more alphanumeric characters and *underscore*, it is possible to represent each message as a vector $\vec{x} = \langle x_1, ..., x_n \rangle$, where $x_1, ..., x_n$ are attribute values $X_1, ..., X_n$ related to the terms $t_1, ..., t_n$. The attributes are integer values obtained from the term frequency (*TF*), representing how often each term occurs in the message.

In addition, any preprocessing was performed, such as *stop words* removal or *stemming*, since some research results indicate that such techniques tend to hurt the performance of the spam classifiers [6].

The classification methods evaluated in this work are listed in Table II. Such methods were selected because most of them have been considered as the best machine learning and data mining techniques currently available [18], [19].

Table II. CLASSIFICATION METHODS EVALUATED IN THIS WORK.

Classification techniques	
CART	Decision trees
k-NN	K-nearest neighbors
LR	Logistic regression
NB-B	Bernoulli Naïve Bayes
NB-G	Gaussian Naïve Bayes
NB-M	Multinomial Naïve Bayes
RF	Random forests
SVM-L	Support vector machines with linear kernel
SVM-P	Support vector machines with polynomial kernel
SVM-R	Support vector machines with Gaussian kernel

As spam filtering is sensitive to chronological order, cross-validation is not recommended to address the algorithms performances [17]. Therefore, we have performed a stratified holdout validation, with the first 70% messages for training and the remainder 30% for testing, so that the earliest messages were used to train the algorithms and the newest ones to test them.

To compare the algorithms performances, we have employed well-known measures used in spam filtering, such as: accuracy rate (Acc), spam caught rate (SC), blocked ham rate (BH), F-measure and Matthews correlation coefficient (MCC) [6].

Blocked ham rate and *MCC* are commonly used to identify methods that achieve a high accuracy at the cost of blocking many legitimate messages. Since a blocked ham is more harmful than a non-caught spam, both measures are important to check the balance between the amount of correctly spam caught and incorrectly blocked ham.

We have employed *grid search* to found the best configuration settings for each evaluated method. For this, we have used 10-fold cross-validation in the training set, so the best configuration could be found without overfitting the model. The following parameters were analyzed using $10^i, -5 \leq i \leq 5$ as the search range: α for NB-B and NB-M; C for SVM-L and LR; and C and γ for SVM-R, SVM-P. The number of trees of RF technique was fitted with the search range 10 to 100, with step size of 10. The best values found for each dataset are reported in Table III.

Table III. PARAMETER SETTINGS ACHIEVED BY GRID SEARCH.

		Dataset				
Method	Parameter	Psy	KatyPerry	LMFAO	Eminem	Shakira
LR	C	10	10^5	10^2	10^2	10^2
NB-B	α	1	10	10^{-2}	10^{-3}	10^{-5}
NB-M	α	1	10	10^{-5}	10^{-1}	10^{-2}
RF	# trees	80	40	90	30	30
SVM-L	C	10^{-1}	10^{-1}	1	10^{-1}	1
SVM-P	C	10^{-5}	10^{-4}	10^{-3}	10^{-5}	10^{-4}
	γ	10	10	10	10^2	10
SVM-R	C	1	10^3	10^3	10^3	10^2
	γ	10^{-2}	10^{-2}	10^{-3}	10^{-4}	10^{-2}

The classification methods, grid search and experiments, were implemented and performed in `Python 2.7.6` using `scikit-learn v.0.16.1` library [20]. The k-NN method was tested with values of k equal to 1, 3 and 5. All parameters not set by grid search were kept with their default values. For methods with random initialization, as Decision trees and Random forests, the seed of random number generator was set to be equal to 0, for reproducibility purpose.

IV. RESULTS

Table IV presents the results achieved by the classifiers over each dataset. The results are sorted by MCC and bold values indicate the highest score for each performance measure.

The results indicate that evaluated classification methods are suitable to automatically filter comment spam posted on YouTube. The most of evaluated methods were able to achieve accuracy rates higher than 90% and blocked ham rates lower than 5%. For instance, the results achieved by RF in Katy Perry's dataset and NB-B in LMFAO's dataset show those methods were able to caught the most of comment spam with zero blocked ham.

False-negatives, or blocked ham, are considered harmful in this kind of application. The methods NB-M and NB-G presented highest rates of blocked ham, mainly in KatyPerry, Eminem and Shakira datasets. In some cases, those methods blocked 1 legitimate comment of 5 analyzed comments. This behavior is unacceptable in automatic spam filtering task.

To ensure the results were not obtained by chance, we have performed a statistical analysis using the non-parametric Friedman test [21] by carefully following the methodology described in Japkowicz and Shah [22]. The Friedman test checks if the null hypothesis, which states there is no difference between the results, can be rejected based on ranking position of each classifier over each dataset. The ranking was built using MCC rates, where the method with the highest MCC for a certain dataset is ranked as 1, and the method with the lowest MCC for the same dataset is ranked as n, where n is the number of classification methods.

Figure 2 shows the ranking of evaluated classification methods, representing the lowest, average and the highest ranking position achieved individually among all five datasets. It is possible to note that NB-M has the largest range among all methods, which means it has achieved very good and very bad results, not being consistently the best or the worst at all. However, k-NN consistently presented the worst performance

Table IV. RESULTS ACHIEVED BY CLASSIFIERS OVER EACH DATASET.

Psy

Methods	Acc (%)	SC (%)	BH (%)	F-measure	MCC
SVM-R	**96.23**	**94.34**	1.89	**0.962**	**0.925**
SVM-L	95.28	90.57	**0.00**	0.950	0.910
LR	95.28	92.45	1.89	0.951	0.907
CART	94.34	90.57	1.89	0.941	0.889
NB-B	93.40	88.68	1.89	0.931	0.872
NB-M	93.40	90.57	3.77	0.932	0.869
RF	91.51	84.91	1.89	0.909	0.838
3-NN	91.51	90.57	7.55	0.914	0.830
SVM-P	90.57	83.02	1.89	0.898	0.821
1-NN	89.62	86.79	7.55	0.893	0.794
5-NN	88.68	81.13	3.77	0.878	0.783
NB-G	88.68	83.02	5.66	0.880	0.779

Katty Perry

Methods	Acc (%)	SC (%)	BH (%)	F-measure	MCC
RF	**94.34**	88.68	**0.00**	**0.940**	**0.893**
LR	92.45	86.79	1.89	0.920	0.855
NB-B	91.51	83.02	**0.00**	0.907	0.842
SVM-L	91.51	84.91	1.89	0.909	0.838
SVM-R	91.51	84.91	1.89	0.909	0.838
NB-M	91.51	**92.45**	9.43	0.916	0.830
NB-G	85.85	86.79	15.09	0.860	0.717
3-NN	83.96	69.81	1.89	0.813	0.708
CART	84.91	79.25	9.43	0.840	0.703
1-NN	79.25	62.26	3.77	0.750	0.622
SVM-P	78.30	58.49	1.89	0.729	0.616
5-NN	77.36	54.72	**0.00**	0.707	0.614

LMFAO

Methods	Acc (%)	SC (%)	BH (%)	F-measure	MCC
NB-B	**97.73**	95.77	**0.00**	0.978	**0.955**
SVM-L	**97.73**	95.77	**0.00**	0.978	**0.955**
CART	**97.73**	**97.18**	1.64	**0.979**	0.954
SVM-R	96.97	95.77	1.64	0.971	0.940
LR	96.21	94.37	1.64	0.964	0.925
NB-G	93.94	88.73	**0.00**	0.940	0.886
RF	92.42	88.73	3.28	0.926	0.852
NB-M	90.91	95.77	14.75	0.919	0.819
1-NN	89.39	87.32	8.20	0.899	0.789
5-NN	88.64	81.69	3.28	0.885	0.785
SVM-P	87.12	78.87	3.28	0.868	0.759
3-NN	87.12	83.10	8.20	0.874	0.747

Eminem

Methods	Acc (%)	SC (%)	BH (%)	F-measure	MCC
CART	**97.78**	97.30	1.64	**0.980**	**0.955**
SVM-R	97.04	95.95	1.64	0.973	0.941
LR	97.04	95.95	1.64	0.973	0.941
SVM-L	96.30	94.59	1.64	0.966	0.926
RF	95.56	95.95	4.92	0.959	0.910
SVM-P	94.81	90.54	**0.00**	0.950	0.901
NB-B	94.81	97.30	8.20	0.954	0.896
5-NN	91.11	83.78	**0.00**	0.912	0.837
NB-G	91.11	**100.00**	19.67	0.925	0.831
3-NN	90.37	83.78	1.64	0.905	0.819
NB-M	88.89	97.30	21.31	0.906	0.783
1-NN	87.41	86.49	11.48	0.883	0.748

Shakira

Methods	Acc (%)	SC (%)	BH (%)	F-measure	MCC
NB-M	**96.43**	**100.00**	6.78	**0.964**	**0.931**
CART	**96.43**	96.23	3.39	0.962	0.928
NB-B	95.54	**100.00**	8.47	0.955	0.915
RF	95.54	90.57	**0.00**	0.950	0.914
NB-G	93.75	**100.00**	11.86	0.938	0.882
SVM-L	93.75	90.57	3.39	0.932	0.876
SVM-R	93.75	90.57	3.39	0.932	0.876
LR	93.75	90.57	3.39	0.932	0.876
3-NN	90.18	79.25	**0.00**	0.884	0.817
1-NN	90.18	79.25	**0.00**	0.884	0.817
SVM-P	86.61	71.70	**0.00**	0.835	0.756
5-NN	81.25	60.38	**0.00**	0.753	0.667

for all datasets. In the bottom of chart, with the best ranking positions, are methods CART, LR, NB-B, RF, SVM-L and SVM-R.

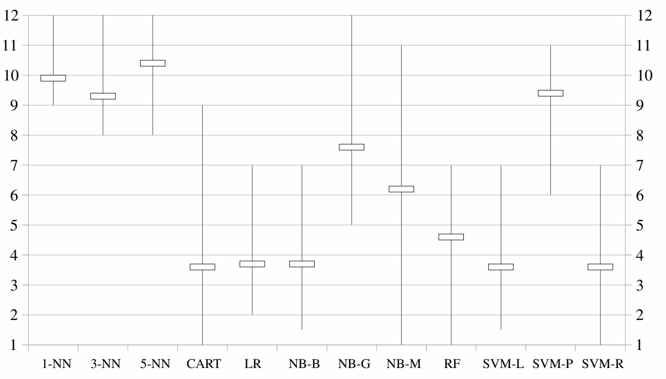


Figure 2. Ranking of classification methods according to their performances achieved by MCC over datasets. The method with the highest MCC is ranked as 1, while the method with the lowest MCC is ranked as n, where n is the number of classification methods.

According to Friedman test, the null-hypothesis can be rejected with a 99.9% confidence level, and therefore there are differences among the performances achieved by evaluated classification methods. However, only a post-hoc test is able to evince which methods are different from each other.

The post-hoc test of Nemenyi [23] was employed to compare the methods pairwise, showing that CART, LR, NB-B, RF, SVM-L and SVM-R are statistically equivalent with 99.9% confidence level, and therefore, they have performed statistically better than any other evaluated techniques.

V. TUBESPAM

Based on results achieved in the reported experiments, we have developed an online tool called TubeSpam[9], that filters undesired YouTube comments right on the fly.

Figure 3 shows the application's home page, in which it is possible to select a video among two lists: the ones with comments already classified on TubeSpam and the most popular ones on YouTube. It is also possible to choose an arbitrary video using its YouTube video *ID*.

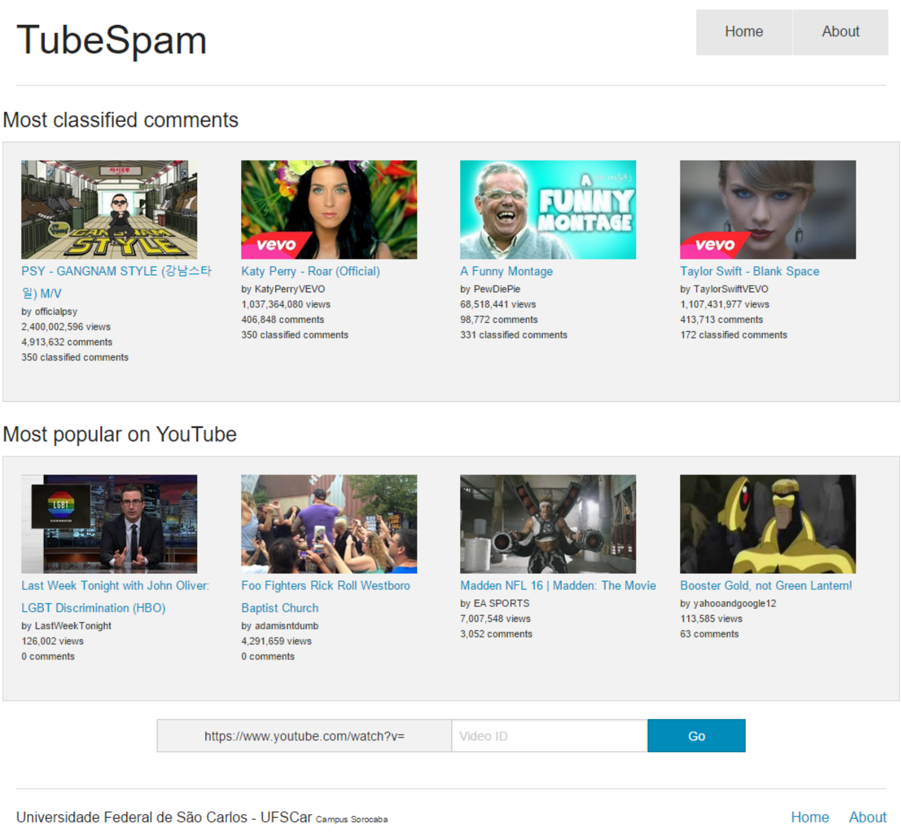
There is a general default classifier to be initially used for any new video. Once the user informs which comments were wrong classified, it enables an option to use those corrections in order to generate a new specific classifier for that video and all videos from its channel. This new classifier also may incrementally learn from corrections made in any video of its channel.

All classifiers are generated using the Bernoulli Naïve Bayes technique. This specific algorithm was chosen because in our experiments it offered a good balance between robustness and computational effort. In addition, this technique offers incremental training, which is suitable for an online tool such as TubeSpam.

Figure 4 shows an example of TubeSpam in action, with comment spam being automatically filtered.

Although it is still in beta stage, the tool is achieving good results on classification process, with accuracy rates around 95% in the training phase.

[9] *TubeSpam*. Available at http://lasid.sor.ufscar.br/ml-tools/.



TubeSpam

Most classified comments

PSY - GANGNAM STYLE (강남스타일) M/V
by officialpsy
2,400,002,596 views
4,913,632 comments
350 classified comments

Katy Perry - Roar (Official)
by KatyPerryVEVO
1,037,364,080 views
406,848 comments
350 classified comments

A Funny Montage
by PewDiePie
68,518,441 views
98,772 comments
331 classified comments

Taylor Swift - Blank Space
by TaylorSwiftVEVO
1,107,431,977 views
413,713 comments
172 classified comments

Most popular on YouTube

Last Week Tonight with John Oliver: LGBT Discrimination (HBO)
by LastWeekTonight
126,002 views
0 comments

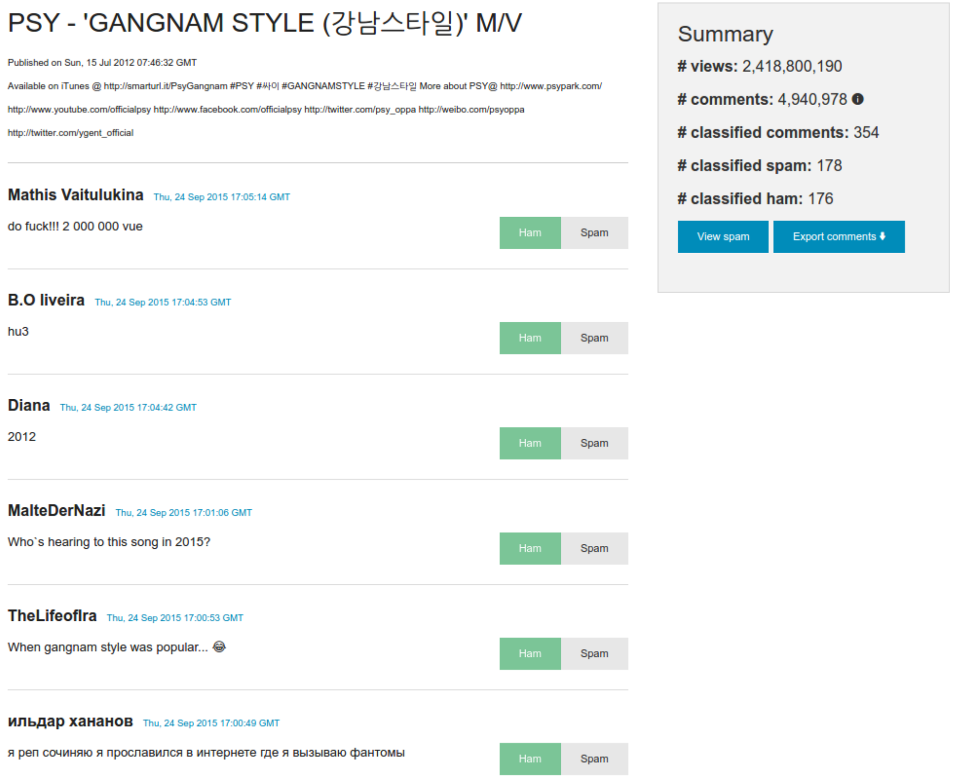
Foo Fighters Rick Roll Westboro Baptist Church
by adamisntdumb
4,291,659 views
0 comments

Madden NFL 16 | Madden: The Movie
by EA SPORTS
7,007,548 views
3,052 comments

Booster Gold, not Green Lantern!
by yahooandgoogle12
113,585 views
63 comments

Universidade Federal de São Carlos - UFSCar Campus Sorocaba

Figure 3. TubeSpam home page. The user can pick a pre-selected video or inform a specific video ID in the bottom of the page.



PSY - 'GANGNAM STYLE (강남스타일)' M/V

Published on Sun, 15 Jul 2012 07:46:32 GMT

Available on iTunes @ http://smarturl.it/PsyGangnam #PSY #싸이 #GANGNAMSTYLE #강남스타일 More about PSY@ http://www.psypark.com/ http://www.youtube.com/officialpsy http://www.facebook.com/officialpsy http://twitter.com/psy_oppa http://weibo.com/psyoppa http://twitter.com/ygent_official

Mathis Vaitulukina Thu, 24 Sep 2015 17:05:14 GMT
do fuck!!! 2 000 000 vue

B.O liveira Thu, 24 Sep 2015 17:04:53 GMT
hu3

Diana Thu, 24 Sep 2015 17:04:42 GMT
2012

MalteDerNazi Thu, 24 Sep 2015 17:01:06 GMT
Who`s hearing to this song in 2015?

TheLifeofIra Thu, 24 Sep 2015 17:00:53 GMT
When gangnam style was popular... 😂

ильдар хананов Thu, 24 Sep 2015 17:00:49 GMT
я реп сочиняю я прославился в интернете где я вызываю фантомы

Summary

# views: 2,418,800,190
# comments: 4,940,978
# classified comments: 354
# classified spam: 178
# classified ham: 176

Figure 4. TubeSpam in action. It has automatically filtered comment spam posted on Psy – Gangnam Style video.

VI. Conclusion and future work

YouTube is a video content publication platform with social network features. Recently, due its popularity and a brand new monetization system, the platform was flooded by undesired content, usually made of low quality information known as *spam*, posted by users looking for self-promotion or to disseminate malicious links.

Automatic spam filtering on YouTube comments is still an unexplored field, evinced by lack of available tools for comment moderation. In this way, owners of famous channels have disabled comments in their videos, choosing other platforms for keep in touch with their audience.

Given this scenario, the main goals of this work were to find promising methods and settings that could be used to assist the detection of undesired comments on YouTube; to offer new public datasets and good baseline results for future comparisons; and to provide an online tool able to automatically detect comment spam posted in YouTube videos.

Firstly, five datasets were collected using public and non-encoded data extracted directly from YouTube, whose were labeled and employed to evaluate several established classification methods.

The results have indicated that the most of evaluated classification methods are indicated for filtering comment spam on YouTube. In fact, the most of them were able to achieve accuracy rates higher than 90% with low or even zero blocked ham rates. The Friedman test assured that results were not obtained by chance, then Nemenyi post-hoc test was employed to compare the methods pairwise. The post-hoc showed CART, LR, NB-B, RF, SVM-L and SVM-R present performances statistically equivalent, with a 99.9% confidence level.

For future work, since there was not just one method that achieved the best result for every single dataset, we can suppose an ensemble of classification methods can lead to better performance than single classifiers. We also aim to employ text normalization techniques and semantic indexing to preprocessing the messages, since they are very short and rife with idioms, slangs, symbols, emoticons and abbreviations. Regarding the TubeSpam tool, we intend to develop web browser plugins to filter spam directly from YouTube.

Acknowledgment

The authors thank to CAPES and FAPESP grants #2014/01237-7 for the financial support.

References

[1] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *Proc. of the 1st AIRWeb*, Chiba, Japan, 2005, pp. 1–6.

[2] R. M. Silva, T. A. de Almeida, and A. Yamakami, "Artificial Neural Networks for Content-based Web Spam Detection," in *Proc. of the 2012 ICAI*, Las Vegas, NV, EUA, 2012, pp. 1–7.

[3] C. Romero, M. Valdez, and A. Alanis, "A comparative study of machine learning techniques in blog comments spam filtering," in *Proc. of the 6th WCCI*, Barcelona, Spain, 2010, pp. 63–69.

[4] T. C. Alberto and T. A. Almeida, "Aprendizado de máquina aplicado na detecção automática de comentários indesejados," in *Anais do X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'13)*, Fortaleza, Brazil, 2013.

[5] Z. Li and H. Shen, "Soap: A social network aided personalized and effective spam filter to clean your e-mail box," in *INFOCOM, 2011 Proceedings IEEE*, April 2011, pp. 1835–1843.

[6] T. Almeida, J. Almeida, and A. Yamakami, "Spam filtering: How the dimensionality reduction affects the accuracy of naive bayes classifiers," *Journal of Internet Services and Applications, JISA'11*, vol. 1, no. 3, pp. 183–200, 2011.

[7] J. M. Gómez Hidalgo, T. Almeida, and A. Yamakami, "On the Validity of a New SMS Spam Collection," in *Proc. of the 11st ICMLA*, vol. 2, Miami, FL, EUA, 2012, pp. 240–245.

[8] T. P. Silva, I. Santos, T. A. Almeida, and J. M. Gómez Hidalgo, "Normalização Textual e Indexação Semântica Aplicadas na Filtragem de SMS Spam," in *Proc. of the 11st ENIAC*, São Carlos, Brazil, 2014, pp. 1–6.

[9] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," in *Proc. of 3rd WWE*, Edinburgh, UK, 2006, pp. 1–8.

[10] V. Chaudhary and A. Sureka, "Contextual feature based one-class classifier approach for detecting video response spam on youtube," in *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on*, July 2013, pp. 195–204.

[11] R. Chowdury, M. Monsur Adnan, G. Mahmud, and R. Rahman, "A data mining based spam detection system for youtube," in *Digital Information Management (ICDIM), 2013 Eighth International Conference on*, Sept 2013, pp. 373–378.

[12] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves, "Detecting spammers and content promoters in online video social networks," in *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, 2009, pp. 620–627.

[13] J. M. Campanha, J. V. Lochter, and T. A. Almeida, "Detecção automática de spammers em redes sociais," in *Anais do XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC'14)*, São Carlos, Brazil, 2014.

[14] T. Almeida and A. Yamakami, "Occam´s razor-based spam filter," *Journal of Internet Services and Applications*, vol. 3, no. 3, pp. 245–253, 2012.

[15] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *The 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS'11)*, Perth, Australia, 2011, pp. 46–54.

[16] A. Severyn, A. Moschitti, O. Uryupina, B. Plank, and K. Filippova, "Opinion mining on youtube," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 1252–1261.

[17] A. Bratko, B. Filipič, G. V. Cormack, T. R. Lynam, and B. Zupan, "Spam filtering using statistical data compression models," *J. Mach. Learn. Res.*, vol. 7, pp. 2673–2698, 2006.

[18] X. Wu, V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *KAIS*, vol. 14, no. 1, pp. 1–37, 2008.

[19] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, Jan. 2014.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, pp. 86–92, 1940.

[22] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms - A Classification Perspective*. Cambridge University Press, 2011.

[23] P. F. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton University, 1963.