

Eddieb Sadat
SYS-623 Data Science and Knowledge Discovery
Professor Feng Liu
5/14/23

Identifying YouTube Spam Comments

Abstract

In many media sharing applications is a feature for users to interact with other viewers by being able to write and reply to comments. One such application is YouTube, which specializes in video uploads and fosters community building by allowing users to comment. In order to protect users from malicious and spam comments, YouTube implements spam detection that flags spam messages and prevents them from being visible. However, there are still some comments that bypass these filters and end up exposing innocent users to malicious links and undesirable comments. In order to improve user experience, it is essential to build a better spam detection model. In this report, I used a dataset that contains YouTube comments and classifies them as Spam or Ham (not spam). I attempted to build a spam detection model using classification models in Python and creating new attributes to use. The final spam detection model had a spam detection accuracy of 91.3%, and was achieved by using Natural Language Processing (NLP) techniques to identify keywords from the Spam and Ham comments.

Introduction

YouTube is one of the biggest media platforms, with over 2 billion monthly active users^[1]. With such a massive number of users, videos and comments being updated every day, it is no surprise that YouTube is unable to detect and prevent all spam messages from being posted. Though it's easy to think that spam comments are easy to ignore, the reality is there is a large number of younger users who are susceptible to clicking harmful links on the promise of

something lucrative. Even beyond protecting users from harmful links, spam comments generally ruin the user experience by flooding a comments section with content completely unrelated to the video. With less happy consumers, there is always the possibility of diminishing active users. Thus, in order for YouTube to remain profitable, they must improve their spam detection systems to give users a better experience. I aimed to build a spam detection model that could possibly improve YouTube's and other media platforms' user experience by more accurately detecting spam messages.

In this report, I document the process of obtaining the dataset to build the model, how I created different attributes to more accurately identify spam comments, the attribute selection process, the results of each model iteration, and finally a discussion about the best model results and recommendations going forward.

Data Understanding

The UCI Machine Learning Repository is a massive archive of datasets that are widely used for machine learning and modeling. One dataset, YouTube Spam Collection_[2], was donated to the repository by a group of researchers who collected and manually identified whether YouTube comments were Spam or Ham_[3]. The download consisted of five separate csv files, each containing comments from five of the most popular music videos as of 2015: Psy, KatyPerry, LMFAO, Eminem, and Shakira (Fig. 1).

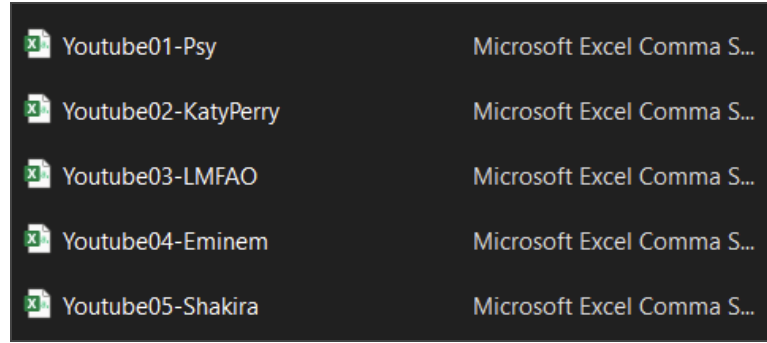


Figure 1 - Five CSV files from Top 5 popular music videos in 2015

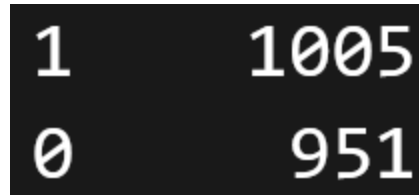
As seen in Figure 2, each csv file contains five columns: COMMENT_ID, AUTHOR, DATE, CONTENT, and CLASS. COMMENT_ID is the unique identifying ID of the comment, AUTHOR is the name of the user who posted the comment, DATE is when the comment was posted, CONTENT is the string containing the comment, and CLASS is the classification of whether the comment is Spam (indicated by a 1) or Ham (indicated by a 0).

COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS
z12pgdhovmrktz	lekanaVEV	2014-07-2	i love this s	1
z13yx345uxepetg	Pyunghhee	2014-07-2	http://ww	1
z12lsjvi3wa5x1vv	Erica Ross	2014-07-2	Hey guys!	1
z13jcjuovxbwfr0g	Aviel Haim	2014-08-0	http://psnl	1

Figure 2 - Five Columns and First few rows from Psy csv file

Data Preparation

To prepare the data for modeling, the contents of each csv were concatenated into one large dataframe. The combined dataframe contained 1,956 rows of data. Next, the COMMENT_ID and DATE columns were removed, as they do not provide useful information for classification modeling. To ensure that the training model remains as unbiased as possible, the distribution of Spam and Ham comments were checked (Fig. 3).



1	1005
0	951

Figure 3 - Number of Spam vs. Ham comments in combined dataframe

Because the number of Spam and Ham comments are very similar, it indicates that this dataset has a healthy distribution of data to continue modeling without any further preparations.

Modeling

To identify the best model, three different classification models from the SciKit Learn python library were tested with various combinations of parameters. The three models tested were SVC (Support Vector Machine), SGD (Stochastic Gradient Descent), and DecisionTreeClassification. With the different model parameters, over 139 different models were tested, each with an equal set-random-state to ensure consistency.

To produce a good model, appropriate attributes need to be developed that are important to properly identify Spam from Ham. In the first iteration of the model, I used very simple attributes as a baseline. Four attributes were created and added to the dataframe (Fig. 4). C_LEN is the number of characters in the comment, A_LEN is the number of characters in the author name, C_SPEC is the number of special characters (<, >, :, ;, [,], /, \, {, }, ! ...) in the comment, and A_SPEC is the number of special characters in the author name.

AUTHOR	CONTENT	CLASS	C_LEN	A_LEN	C_SPEC	A_SPEC
Julius NM	Huh, anyway check out this you[tube] channel: ...	1	56	9	3	0
adam riyati	Hey guys check out my new channel and our firs...	1	166	11	7	0
Evgeny Murashkin	just for test I have to say murdev.com	1	38	16	0	0
ElNino Melendez	me shaking my sexy ass on my channel enjoy ^_^	1	48	15	2	0
GsMega	watch?v=vtaRGgvGtWQ Check this out .	1	39	6	2	0
...
Katie Mettam	I love this song because we sing it at Camp al...	0	58	12	2	0
Sabina Pearson-Smith	I love this song for two reasons: 1.it is abou...	0	93	20	1	1
jeffrey jules	wow	0	3	13	0	0
Aishlin Maciel	Shakira u are so wiredo	0	23	14	0	0
Latin Bosch	Shakira is the best dancer	0	26	11	0	0

Figure 4 - Dataframe with four added attributes

The four attributes were used to produce the 139 models stated above. The best model was produced by the DecisionTreeClassifier, and the results shown in Figures 5 and 6.

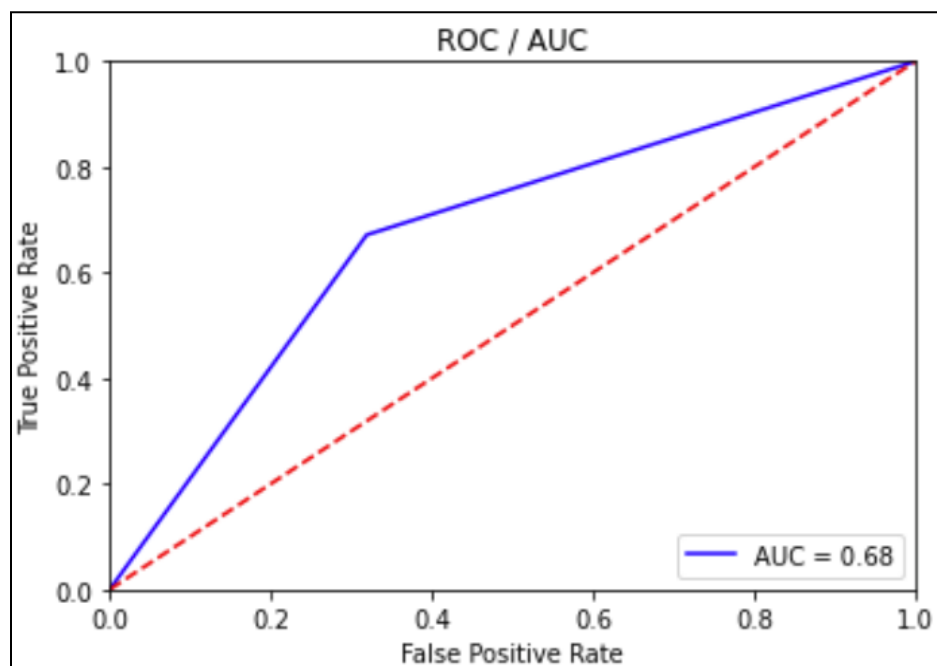


Figure 5 - ROC & AUC of first model

```
[[205  96]
 [ 94 192]]
0.676320272572402
```

Figure 6 - Confusion Matrix of first model

The results show that this first model was only able to accurately identify Spam and Ham comments with about 68% accuracy. To try and improve this model, the importance of each attribute was calculated (Fig. 7).

```
C_LEN    0.581792
A_LEN    0.207746
C_SPEC    0.200967
A_SPEC    0.009494
```

Figure 7 - Importance levels of attributes in first model

These results show that A_SPEC only had an importance of about 0.9%, indicating that this attribute is not necessary for the model. As such, it was removed and new models were generated. The best model was produced by the DecisionTreeClassifier, but the accuracy and AUC saw no change.

Next, new attributes were created by using NLP techniques to identify key words. First, the dataset was split into Ham and Spam comments. Using the nltk Python library, each comment was modified and analyzed to separate and identify each word in the string. The frequency of all unique words from the comments were calculated (Fig. 8).

```
Most common tokens in Spam:Most common tokens in Ham:
[('check', 559),          [('song', 224),
 ('com', 296),            ('love', 145),
 ('please', 246),         ('like', 90),
 ('youtube', 235),        ('views', 87),
 ('subscribe', 229),      ('video', 84),
 ('video', 229),          ('br', 64),
 ('39', 210),             ('best', 57),
 ('channel', 197),        ('katy', 56),
 ('br', 195),             ('2', 55),
 ('like', 160)]           ('billion', 51)]
```

Figure 8 - Top Ten most frequent words of Spam and Ham comments

To produce the best results, I decided to only select words from each list that did not show up in both Spam and Ham. This will ensure that the words selected are only more likely attributed with Ham or Spam. The words selected were check, com, subscribe, please, channel, youtube, views, song, and love. Nine new binary attributes were created to indicate whether each of these words is in the comment. New models were produced with only these nine new attributes. The best result occurred with the SVC model with kernel set to poly and C set to 5. Figures 9 and 10 show the ROC/AUC and confusion matrix.

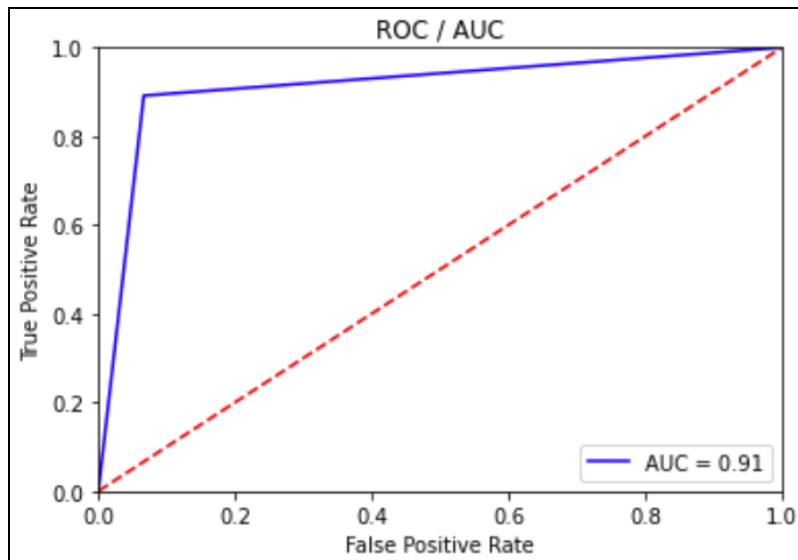


Figure 9 - ROC and AUC of nine attribute model

```
[[281  20]
 [ 31 255]]
0.9131175468483816
```

Figure 10 - Confusion matrix of nine attribute model

Clearly these nine attributes were vastly more successful in producing a model, which was able to correctly classify comments as Spam or Ham approximately 91.3% of the time. To try and improve the model, the importance of each attribute was calculated (Fig. 11).

```
check      0.333252
com        0.224855
subscribe  0.189099
please     0.080526
channel    0.067594
youtube    0.047107
views      0.027215
song       0.019330
love       0.011023
```

Figure 11 - Importance levels of nine NLP attributes

From the results, love, song, and views were removed from the dataframe since they each had important levels less than 3%. New models were then generated, however, none of the results were better than when the model contained all nine attributes, and the best model was actually less accurate than before at only 90%.

Finally, in the last model iteration, the four character count attributes were combined with the nine keyword attributes and new models were generated. The best model was created by SVC with kernel set to linear and C set to 1. The ROC/AUC and confusion matrix are shown in Figures 12 and 13.

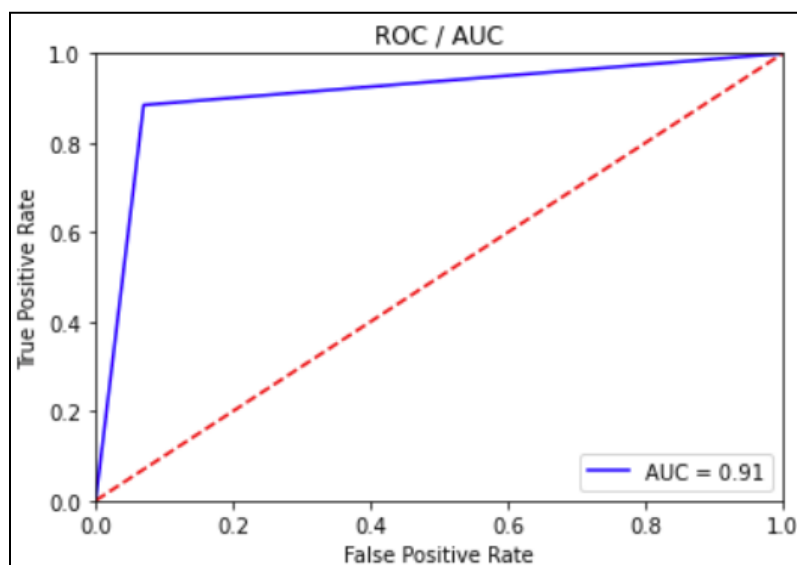


Figure 12 - ROC and AUC of combined attribute model

```
[[280  21]
 [ 33 253]]
0.9080068143100511
```

Figure 13 - Confusion matrix of combined attribute model

This model was still worse than the model that only used the nine attributes with an accuracy of about 90.8%. To try and improve the model, the importance levels of each attribute was calculated (Fig. 14).

C_LEN	0.218954
check	0.196641
C_SPEC	0.127033
subscribe	0.120458
com	0.089619
A_LEN	0.067099
channel	0.052361
please	0.046266
youtube	0.030081
song	0.021453
views	0.017116
love	0.011446
A_SPEC	0.001475

Figure 14 - Importance levels of combined attributes

The last four attributes had an importance level less than 3%, and as such they were removed from the dataframe. New models were produced with the updated attributes, and just as before it resulted in a worse model with an accuracy of about 90.1%.

Discussion

Ultimately, the best model had a classification accuracy of 91.3% and was produced by an SVC model with kernel set to poly and C set to 5 that used all of the identified keywords. Although an accuracy that high may be good enough to be deployed in a real setting, it is important to note that the model is likely very biased towards popular music videos. Because the dataset only contained comments from the top five most popular music videos, the comments may be specifically geared towards that type of content, which is evident by the high frequency

of the word 'song'. This means that the model may become very inaccurate with comments from other types of videos, meaning it should not be deployed until much more testing is completed.

In conclusion, NLP clearly showed superiority in creating a spam detection model, and I would highly recommend further research in improving spam detection by using NLP techniques to identify keywords in comments.

References:

- [1] - <https://buffer.com/library/social-media-sites/>
- [2] - <https://archive.ics.uci.edu/ml/datasets/Youtube+Spam+Collection>
- [3] - <http://dcomp.sor.ufscar.br/talmeida/youtubespamcollection/>