YouTube Comments Spam filter

Introduction

Youtube is one of the most popular video sharing platforms with more than 1 billion users. Users have long been outraged by the overwhelming number of spam messages in the comment section. In 2012 users created a petition asking Youtube to provide tools to deal with undesired content. In 2013, the spam problem worsened as Google overhauled the YouTube comment system to connect it to Google+, which allows users to post links. This attracted more malicious users to self-promote their videos using the platform. This project will build a spam filter to automatically filter spam comments.

Who might care?

Spam comments are annoying for content creators, for the users as well as platform owner. Content creators want to read about how their fans reacts to the video they created. Users want to glance through the comments to connect with other fans as well as with their favorites content creators. Platform owner want to get rid of malicious spam to

make sure their platform is safe to use and provides an enjoyable experience to all end users.

Exploratory Data Analysis

The data is acquired from UCI machine learning repository. The data is collected through YouTube API, extracted from 5 most popular videos on Youtube during the first half of 2015. There are total of 1956 comments in this dataset for the 5 most popular songs from 5 artists: Psy, Katy Perry, LMFAO, Eminem, Shakira. 951 are labeled as hams and 1005 are labeled as spams. There are ~200 comments (~10%) which are missing date information, since the comments appear sequentially in time, we can use either forward fill or backward fill to estimate the time when that comment is made. In this report I used forward fill. The CLASS section is the label: 1 means spam and 0 means ham. Here is a subset of the dataset after cleaning.

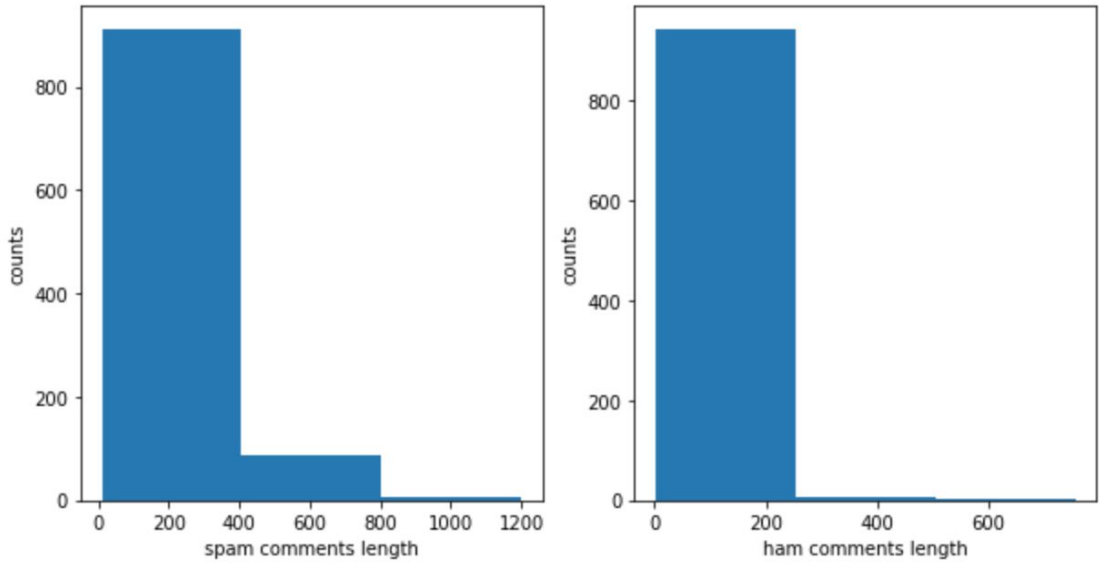
	COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS	song
0	LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU	Julius NM	2013-11-07T06:20:48	Huh, anyway check out this you[tube] channel: ...	1	Psy
1	LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A	adam riyati	2013-11-07T12:37:15	Hey guys check out my new channel and our firs...	1	Psy
2	LZQPQhLyRh9MSZYnf8djyk0gEF9BHDPYrrK-qCczIY8	Evgeny Murashkin	2013-11-08T17:34:21	just for test I have to say murdev.com	1	Psy
3	z13jhp0bxqncu512g22wvzkasxmvvzjaz04	ElNino Melendez	2013-11-09T08:28:43	me shaking my sexy ass on my channel enjoy ^_^	1	Psy
4	z13fwbwp1oujthgqj04chlngpvzmtt3r3dw	GsMega	2013-11-10T16:05:38	watch?v=vtaRGgvGtWQ Check this out .	1	Psy

Some interesting questions I explored on the dataset:

- Is the average length of comment different b/t spam and ham group?
- Does spam more likely to come from foreign accounts?
- Does spam group tend to have more capital letters?
- Are spam comments more likely to have a URL in them?
- Does the spam comments have any correlation with time?
- What are most common words used in spam and ham?
- Can we decompose spam/ham into topics and explore using LDA?
- Can we find different types of spam?
- Can we visually inspect spam/ham?

Average comment length:

Below is the comment length distribution among spam and ham group. From the summary statistics it's seen that the average comment length is quite different. T-test shows that the difference in the average comment length is significant among two groups.



CLASS	mean	std	min	max
0	49.827550	56.526731	2	755
1	137.769154	159.459172	10	1200

Foreign accounts:

Out of all 1956 comments made, there are 1793 unique user names in this dataset. There is a small fraction of users who make multiple comments but all of users who created spam only comment once. Some spam accounts are from foreign countries. The percentage of

foreign accounts in spam/ham group is comparable. T-test doesn't show significant difference among the two groups.

CLASS	user_isEnglish	
0	True	901
	False	50
1	True	956
	False	49

Capital letters:

The number of capital letters adjusted to the comment length is summarized as below (i.e. the percentage of capital letters in a comment). Although the average percentage of capital letters are very close among two groups, t-test shows(p=0.016) the difference is significant.

CLASS	mean	std	min	max
0	0.090289	0.161128	0.0	1.000000
1	0.108451	0.173301	0.0	0.919355

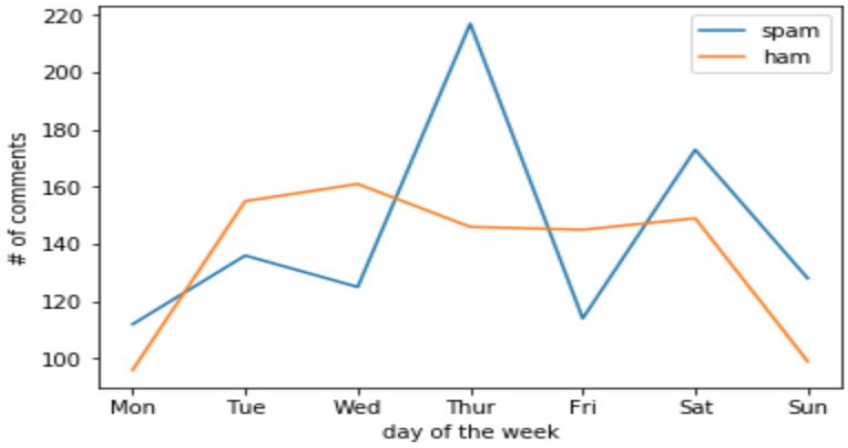
Do spams tend to have URLs?

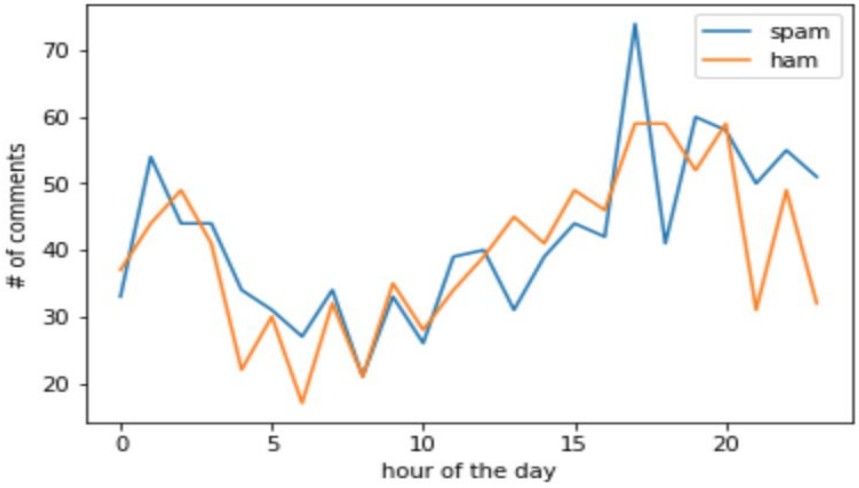
YES! About 14% of the comments contains URL, and most URLs are found in spam comments. The percentage of spam comments containing URL is 5.8 times more than ham comments containing URL. (spam: 85.4% and ham: 14.6%)

Correlation with time:

We can plot the number of comments vs the week of the day or hours to see the pattern. These summary graphs show interesting user behaviour. As expected, users tend to make more comments in late afternoon throughout the night. You can see most spam and ham comments are made b/t 5pm to 1 or 2am. Less comments are made on Sunday and Monday. Probably due to people being busy starting their new week (prep food for kids, run errands etc) and spend less time online. While ham comments are made evenly

throughout the rest of week (Tuesday to Saturday), spam comments spike on Thursday and show some zigzag behavior.





The most common words in spam and ham:

To find most frequent words among two groups which conveys meanings, I used some preprocessing technique on the text input, these are summarized as below:

1. Split into tokens.
2. Convert to lowercase.
3. Remove punctuation from each token.
4. Filter out remaining tokens that are not alphabetic.
5. Filter out tokens that are stop words.

The 10 most frequent words in spam are:
'subscrib', 'video', 'http', 'amp', 'like', 'br', 'check', 'pleas', 'youtub', 'channel'

The 10 most frequent words in ham are:
'love', 'perri', 'billion', 'video', 'song', 'like', 'br', 'best', 'kati', 'view'

We can see that the most frequent words in spam group relate to self-promoting videos with keywords like http, check and subscribe. The most frequent words in ham group however are more related to the song, some keywords relate to the song's artist like kati perri. Some keywords describe the general feelings towards the song such as best, like, love.

Topic analysis LDA:

Using Latent Dirichlet Allocation (LDA) we can model the topics the spam comments and ham comments are drawn from and the most important words within each topic.

Here I select the number of topics to be 5, the topics and the 10 most important words within the topic is shown:

In spam group:

Topics generated in the spam group relate to self-promoting videos such as: check and subscribe channel, go watch, http, etc. It also populates the words that are shared among ham group like kati (the artist) and awesome song (general feeling).

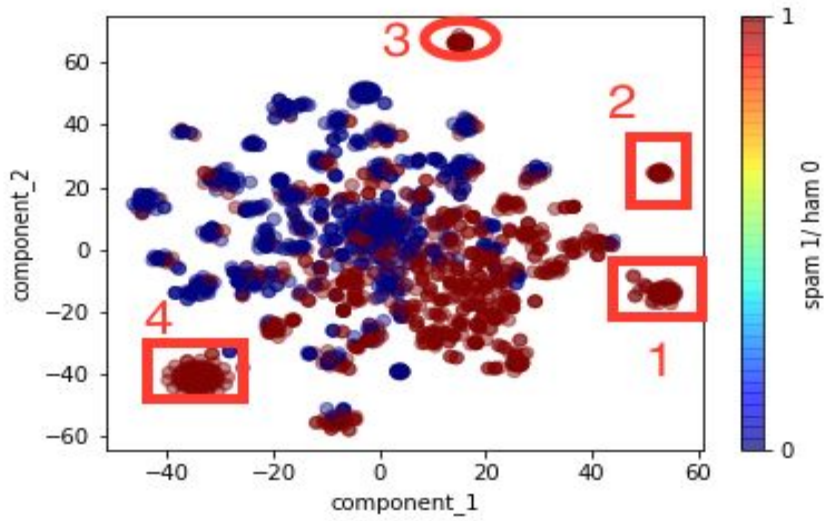
```
Topic #0:
check video youtub subscrib channel pleas guy go watch thank

Topic #1:
make new money work share http month home visit websit

Topic #2:
br quot remix song rihanna check like subscrib hit million

Topic #3:
http amp gt like pleas lt kati free awesom song

Topic #4:
music pleas like thank comment u would hey check get
```

In ham group:

In general, the topics in ham group are more related to the content of the songs, for example it picks up artist names: katy perry, psy, gangnam style, shakira, eminem. It also includes more words which describe feelings towards the songs, such as like, awesome, wow, love song, best, beautiful, nice.

```
Topic #0:
kati song perri music still lt eminem fuck like listen

Topic #1:
view video like billion song get watch peopl awesom old

Topic #2:
year shuffl time style psi like gangnam would lol back

Topic #3:
good wow go girl megan fox first comment part shit

Topic #4:
love song br shakira best beauti waka make much nice
```

In addition, a dimension reduction can be performed on spam/ham and visualize them on a 2D graph. Here I used t-SNE. We can see some interesting clusters of spam which clearly stands out (labeled 1,2,3,4). A closer look at the comments inside cluster 1/2/3/4 tell us different type of spam messages. All spam comments in cluster 1 contain URLs. All spam comments in cluster 2 contain the exact same comment: “You guys should check out this EXTRAORDINARY website called FIREPA.COM ….”. Cluster 3 contains the exact same comment: “Check out this playlist on YouTube:”. Cluster 4 contains the exact same comment: “Check out this video on YouTube:” The probability of making the exact same comment and use exact same capital words is pretty low. If we were to have more user account data, we can trace back these accounts and flag them as suspicious malicious account and use it as additional feature into our machine learning model.



To conclude, in this section, we've looked several features that are different among spam and ham group, such as length of the comment, percentage of capital words, etc. We've also looked at commonly used words and general theme of spam vs ham group and a visual exploration of different types of spam comments. Next we can build machine learning models to filter spam.

Machine Learning Models

The dataset is split into training and test data set. Model hyperparameters tuning is performed using grid search on training set, and test set is evaluated at the last. The comments is fed into Tfidf vectorizer to extract features. It's found that stem and stop word removal hurts the performance on naive bayes thus no preprocessing is performed on the dataset.

Model evaluation: In this application, we can imagine that users will be outraged if their comments are classified as spam. On the other hand, spam message misclassified as ham is not as serious and given that the volume of spam message is small, users can still have an enjoyable experience. Thus we can pick F-beta score with beta=0.5 as our ultimate metric for model selection to place a little more emphasis on precision than recall. In practice, we can have online experiment to quantify the cost of spam vs ham, for example we can test how false positive vs false negative affect user watch time. We can then define a model evaluation metric that takes into account of the cost of false positive vs false

negative. In this project, I used AUC score to grid search model parameters. The benefit of using AUC score is that it's independent of threshold. Since we care more about precision than recall, we can then tune threshold to get a good model that optimizes our F-beta score - giving more priority to precision without overly sacrificing recall. We will then choose the model that achieves the best F-beta score via this methodology.

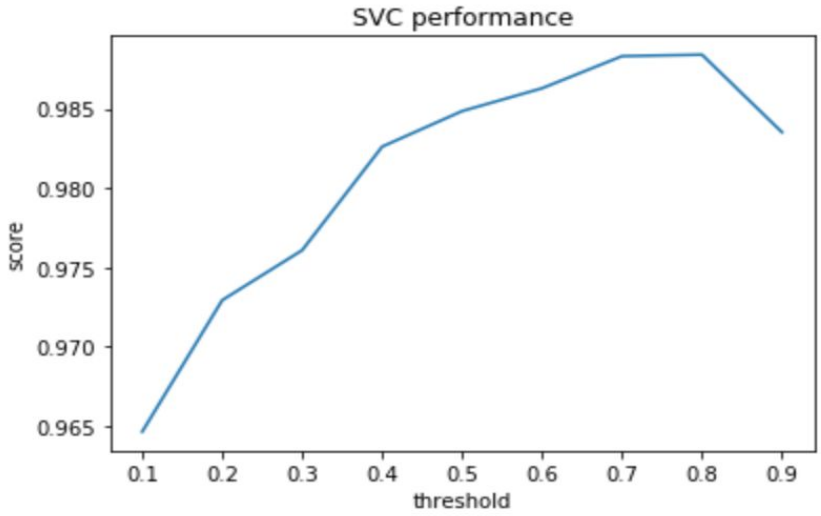
Here is the summary of different model performance after applying this method to each.

Models	F-beta score on training data	F-beta on test data	Precision on test data	Recall on test data	Accuracy on test data
Multinomial NB	0.974	0.954	0.961	0.928	0.947
Random Forest	0.999	0.962	0.973	0.920	0.949
Logistic Regression	0.995	0.959	0.962	0.949	0.957
SVM	0.988	0.965	0.973	0.932	0.955

Seen from the above summary table, SVM has slightly outperformed the other models, with the highest F-beta score at 0.965. It achieves high precision while maintaining good recall rate.

Models	Model parameter name	Model parameter values	Threshold
MultinomialNB	alpha	1	0.6
Random Forest	N_estimators, min_sample_leaf	200,1	0.5
Logistic Regression	C	10	0.5
SVM	'Kernel' , 'C'	'linear' , '1'	0.7

As an example of how to choose the threshold, I graph the F-beta score against threshold and pick the maximum F-beta. Here is the graph for SVM. F-beta peaks at threshold = 0.7, thus 0.7 is picked.



We also see from the summary table that F-beta score is higher on training data than on test data. This indicates overfitting. Some of the methods we can use to address overfitting are:

#1: dimension reduction. We can do a PCA first to reduce the dimension of our features before feeding into machine learning models.

#2: regularization. For example, in logistic regression, we can use ridge regularization which penalize large coefficients.

#3: In random forest, we can reduce overfitting by limit the depth of a tree.

Here lists some examples of misclassified hams and spams that's common in 2-3 models listed above.

Hams misclassified as spams:

Example #1: " If you are a person that loves real music you should listen to "Cruz Supat"
He is awesome as fuck!!! Just as eminem used to be."

Example #2: "Loves it"

Example #3:
“<ahref="http://www.youtube.com/watch?v=KQ6zr6kCPj8&t=2m19s">2:19 best part”

Spams misclassified as hams:

Example #1: “like if ur watchin in
9000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000”

Example #2: “+447935454150 lovely girl talk to me xxx”

Example #3: “I'm A SUBSCRIBER”

Example #4: “+447935454150 lovely girl talk to me xxx”

Future Improvements

To improve our model, we could hand picked features to tune our model. Here are some ideas for enhancement we can do in the future.

#1: From EDA section, we’ve seen that four features (length of the comment, percentage of capital words, whether it contains URL, whether it’s sent on Thursday) are significantly different among spam and ham. We can use these as additional features to feed into our model.

#2: Examining commonly examples of ham misclassifying as spam in last section, we can make additional features.

For example, “Loves it” is misclassified as spam. We can do an analysis to see if a short comment (such as comment length <=3), is it more likely to be spam or ham? If there is a significant difference then we can make a feature based whether or not a comment is less than 3 characters.

Another example of ham misclassified as spam,

“<ahref="http://www.youtube.com/watch?v=KQ6zr6kCPj8&t=2m19s">2:19 best part”. Youtube comments are short, one feature would mislead the prediction. URL is a strong indicator to be spam thus above comment is easily be classified as spam. However we can check whether or not the URL is linked back to the current video, if so, then it’s more likely to be ham.

Some examples of spam misclassified as ham:

“+447935454150 lovely girl talk to me xxx” and “+447935454150 lovely girl talk to me xxx”.

We can make additional feature whether or not a comment contains phone numbers to help make the model better.

#3 In addition, in an online platform, we would expect a lot of typos. This is not good for bag of words model. For example, if we encounter misspelling in our test data then we won’t find it in our library. We can add a preprocessing step to correct all misspelling before creating bag of words.

#4 For this data set, we don’t have much information about user account. In addition, most user names only make one comment, if we had user account information and had more training sample where we have more samples each account user make, we can learn malicious account.

Due to time constraint, I only tried idea #1 from above and no significant improvement for model accuracy is observed. In case you are interested how the model performs, here is the summary table.

Model name- concat features	F-beta on training data	F-beta on test data	Precision on test data	Recall on test data	Accuracy on test data
SVM	0.987	0.959	0.965	0.936	0.953

Model name- concat features	Model best parameter	Threshold
SVM	'kernel': 'linear', 'C': 1	0.8

To conclude this section, several models have been presented with accuracy on test dataset ~0.96 and F-beta score ~0.97 and further improvement ideas are discussed in detail.

Conclusion

In this report, a spam filter for Youtube comments are presented. The best model in this report is SVM, which can achieve a F-beta score of 0.965 and accuracy of 0.955. SVM is good for small data size (several thousands), however if we were to train on a larger data size, random forest, logistic regression and naive bayes would be better choices. For future work, to make this model work better, we can learn malicious account, correct spelling as preprocessing before creating bag of words, hand pick additional features such as whether or not the comment contains phone numbers, whether or not the comment length is extremely small, whether or not the URL points back to the current video itself etc. Some spam/ham comments are hard to distinguish, however there are examples we can hand pick features that is very different among spam/ham.