Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 7, No. 4, Desember 2014 **ISSN 0216 – 0544**

RESTRICTED CONTENT CLASSIFICATION BASED ON VIDEA METADATA AND COMMENTS (CASE STUDY : YOUTUBE.COM)

[a]Stefanus Thobi Sinaga, [a]Masayu Leylia Khodra
[a,b]Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Jl. Ganesha 10 Bandung
E-Mail: s.thobi.sinaga@gmail.com

Abstrak

Klasifikasi konten terbatas merupakan kegiatan memisahkan konten video yang layak untuk seluruh pengguna dari konten yang tidak layak untuk pengguna di bawah umur (<18 tahun). Pada situs Youtube, proses klasifikasi konten terbatas dilakukan secara manual oleh karyawan berdasarkan laporan yang dikirimkan oleh komunitas pengguna. Pada penelitian ini dirancang sebuah sistem klasifikasi konten terbatas secara otomatis yang dapat melakukan klasifikasi terhadap video Youtube berdasarkan teks metadata (judul, deskripsi) dan komentar dari video tersebut. Sistem tersebut memanfaatkan model klasifikasi hasil eksperimen terhadap *dataset* video Youtube yang telah dikumpulkan. Judul dan deskripsi video dipilih sebagai atribut klasifikasi karena mengandung informasi utama yang ditulis oleh penggunggah terkait video yang diunggah. Sedangkan komentar dipilih sebagai atribut klasifikasi karena dapat dijadikan sumber informasi ketika informasi yang disediakan oleh pengunggah tidak dapat mereprentasikan video yang digunakan. Melalui eksperimen, didapatkan model klasifikasi dengan F-Measure sebesar 83,45%. Model dibangun dengan menggunakan pendekatan leksikal terhadap dataset judul dan deskripsi video (tanpa komentar), Support Vector Machines sebagai algoritma klasifikasi, serta metode *binary* sebagai metode pembobotan fitur. Dengan menggunakan model tersebut, telah dikembangkan sistem klasifikasi konten terbatas berdasarkan teks metadata dan komentar video.

Kata kunci: Klasifikasi, Konten Terbatas, *Support Vector Machines*.

Abstract

Restricted content classification is an activity of labeling video content into two category, which are restricted content that is appropriate for all audiences and non-restricted content that are not appropriate for minor audiences (age < 18). On Youtube, restricted content classification is being processed manually by the expert staffs based on user reports. This research aims to build automatic restricted content classification system which is able to classify Youtube video based on its metadata (title, description) and video comments. This system would use the best model achieved from the experiment on Youtube video dataset. Video title and description are chosen as the classification attribute since they contain the main information about the video provided by the uploader. Meanwhile, video comments are chosen as the other classification attribute under the assumption that they would provide the information necessary when video title and description are not able to give any information related to the video. Our experiment shows that the best classification model with F-Measure of 83.45% is achieved by using lexical feature on dataset built from video title and description (without comments). We employed Support Vector Machines as the classification algorithm and binary as the feature weighting method. In this paper, a restricted content classification system based on metadata and video comments has been built.

Keywords:Classification,Restricted Content, Support Vector Machines.

INTRODUCTION

Internet video sharing is an internet phenomenon that has been exist for many years. Averaging on 300 hours video uploaded per minute and one bilion views per day [1], Youtube.com is the most successful video sharing website available on the internet. All sorts of video can be found on Youtube including video that is not appropriate for minor audiences (age < 18). Those type of videos are called restricted content (Figure 1).

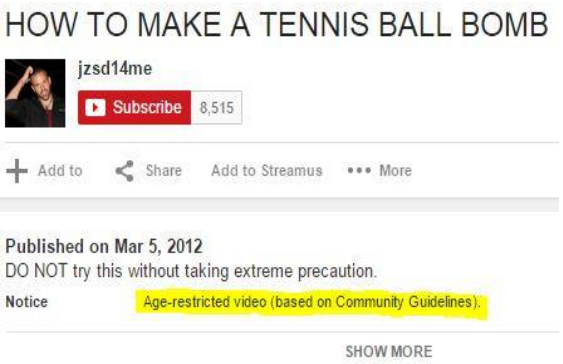

Figure 1. Video restriction information

On Youtube, restricted content is the type of content or videos that contain at least one of vulgar languages, sexual content, gore/violence, and dangerous activity . Currently user can found a lot of resticted content which either has been identified or still remain unidentified. The unidentified restricted content is still remaining because currently it is not possible to evaluate every video uploaded to Youtube. On Youtube, restricted content is evaluated by Youtube's expert staff based on user reports. The limitation of Youtube's expert staff and the enourmous growth of Youtube's video force its restricted content classification system to rely on the user base. This kind of dependency makes Youtube's restricted content classification system less effective because a restricted content can remain unidentified when there is no user that report the content or high number of user reports so that the reported content is still on the evaluation queue and not processed yet. In order to solve those problems faced by Youtube's current restricted content classification system, an automatic restricted content classification system is picked as an alternative solution.

Automatic restricted content classification system is a classification system that use machine learning approach to produce a model that can predict whether a video/content is a restricted content or not by examining the available information. In this research, metadata (title, description) and video comments are selected as the valuable information in classifying a restricted content. Title and description are chosen because they are the main representation of video provided by the uploader. Video comments submitted by users who watch the video is also selected as another valuable information under the assumptions that they can serve alternative information when the title and description can not give any valuable information (bad title, empty description, etc). The model will then use these information to learn the characteristics of each label (restricted and non-restricted) and predict another video label based on its metadata and video comments.

This paper discusses about the process to build the whole automatic restricted content classification system. The idea is to find valuable keyword in video title, description, and comments and use it as the information to determine its correct label. The rest of this paper is organized as follows. The concise information about text classification and related implementation is provided in Section 2. In Section 3, we describe the process of classifying restricted content on youtube using machine learning approcah. The development and implementation of the system will be described on Section 4. For the experiment that had been done, the result and its analysis will be presented in Section 5. And finally, the conclusion of this research is stated in Section 6.

TEXT CLASSIFICATION

Text classification is a problem of classifying a text into its correct labels. Text classification has been used to solve variety of problems such as spam filtering, news grouping, language detection, etc. In classifying text, there are two available approaches, machine learning based classification and rule based classification. In this research, we will used machine learning

approach as an alternative to rule/expert based approach used by Youtube.

Machine learning approach in text classification research has been done before such as [3] which use video title, description, tag, and comments on Youtube video to find its correct category (movie, music, howto, etc). Another research [4] shows the usage of machine learning based text classification on determining whether a web page is an adult website (nsfw) or not.

In short, machine learning based text classification works by learning the distinctive feature owned by each category/label then use it as a knowledge to determine the correct label of another video.

AUTOMATIC RESTRICTED CONTENT CLASSIFICATION SYSTEM

The automatic restricted content classification system that is built for this research has three main components which are data crawler, classifier, and user interface. Data crawler, as mentioned before, is implemented to collect training and testing data for the corpus, classifier will be the best model achieved from the experiment, and user interface the component which handle user inputs and system outputs. The illustration of system architecture can be found on Figure 2.

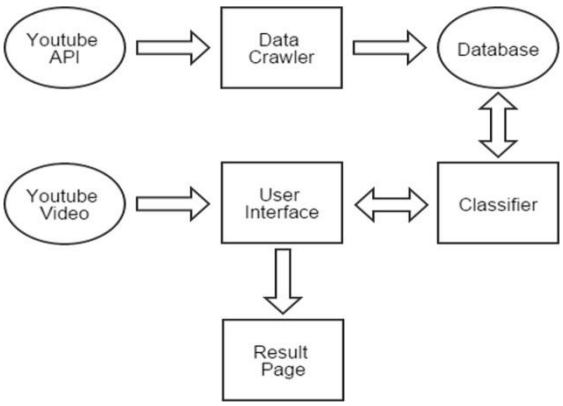

Figure 2. System Architecture

Our data crawler used interface provided by Youtube API to collect video data on Youtube. The video data format/characteristics used by Youtube can be seen on Table 1.

Table 1. Video data characteristics

Type	Characteristic
Title	Text, <= 100 chars. Can't be left empty.
Description	Text, <= 1000 chars. Can be left empty
Comments	Text, <= 1000 chars. There is no minimum or maximum limit of comment on each video.
Restriction rating	The restriction or empty if it is non-restricted

An example of the video instance crawled from the Youtube API is provided on Table 2.

Table 2. Video data instance example

Type	Content
Title	New Action Movies English 2014 Full Movies HD
Description	If you like this channel, please Subscribe for more videos
Comments	- Please add MORE to the list. - AWESOME movie
Restriction rating	scheme='http://gdata.youtube.com/schemas/2007#mediarating' yt:country='all'

Dataset Construction

Using the crawler that has been constructed, 861 instances of video data has been collected. It consist of 485 non-restricted video instances and 376 restricted video instances. Another problem arise as we can not know for sure whether the non-restricted video has been correctly classified / evaluated by Youtube's expert staff before. To solve this problem, re-evaluation of the non-restricted video is needed. Since re-evaluating all of the non-restricted video can be time-consuming, we analyse the restricted video instances to find the threshold needed to pass the assumptions that a video has been evaluated by Youtube's expert staff.

The analysis results shows that 73.13% restricted video instances have user view more than 10.000 (ten thousand) views, 80.31% instances have been uploaded for more than a year, and 66.75% instantces (2 out of 3) both have user view more than 10000 views and

have been uploaded for more than a year. Based on this data, re-evaluation is done to non-restricted video instances that does not meet the threshold requirements (view less than 10000 and has not been uploaded for more than a year). 5 out of 861 non-restricted videos re-labeled as restricted video after the re-evaluation process which leaves the data on 480 non-restricted videos and 381 restricted videos. These instances is then used to construct seven datasets based on the source combination as seen in Table 3.

Table 3. Dataset list

No	Dataset
1	Title
2	Description
3	Comments
4	Title + Description
5	Title + Comments
6	Description + Comments
7	Title + Description + Comments

Dataset Preprocessing

The dataset that has been collected need to be processed before it can be used as a training data. There are four preprocessing methods that need to be aplied to the data. Those are case-folding, punctuation removal, stopwords removal, and stemming. Casefolding is the process of transforming all of the capital letters into its non-capital form. Punctuation removal will remove all the characters beside alphabet (a-z) and numbers (0-9). Stopwords removal will remove all words that are considered as unimportant, meaningless word such as “the”, ”a”, ”an”, etc. Stemming is the process of transforming word into its basic form such as “swimming” into “swim”, “sexier” into “swim”, etc. The stemming algorithm used in this research is Snowball algorithm which is provided by Weka [5].

The example of data preprocessing used in this research can be seen in Table 4.

Feature Extraction and Feature Selection

In this research, lexical-based feature is employed. Lexical approach is chosen because it really fits with the nature of the problems. Restricted content classification is one of many problem that can be solved by finding important keywords for each category. For example words like “sex”, “drugs”, and “violence” more likely to be found on restricted category compared to words like “education”, ”health”, and ”vacation”.

Table 4. Data preprocessing

Stages	Data
Default	AWESOME movie!!! thank you for uploading...great inspiration for the world we’re living in today...
Casefolding	awesome movie!!! thank you for uploading...great inspiration for the world we’re living in today...
Punctuation removal	awesome movie thank you for uploading great inspiration for the world we re living in today
Stopwords removal	awesome movie thank uploading great inspiration world living today
Stemming	awesome movie thank upload great inspire world live today

Syntactic approach will not work well because of the characteristics on metadata and video comments that have lack of structure compared to other text/document such as news document. Semantic approach is suspected to not work well because its complexity tends to not work on problem with simple nature [6].

Feature extraction is conducted by passing the dataset into StringToWordVector provided by Weka [5]. This filter will transform text into a “bag of tokens” that can be weighted using three differents methods (binary, count, tf-idf). Since every words in the documents is transformed into a feature as a token, the number of feature may get too big to be handled by the classifier (in terms of classification time) while there are probabilities that some of those features are not even valuable. Based on that thinking, feature selection is applied to the dataset with expectations that the features will have higher quality in terms of valuable information and the classification time would be reduced significantly. The result of feature weighting can be seen Table 5 while the feature extraction and selection can be seen on Table 6.

Table 5. Feature extraction and weighting result

Process	Data
Input	Title : "Awesome movie 2014" Description : "Upload your movie to http://movie.com, please subscribe"
Feature Extraction	{awesome, movie, upload, movie, please, subscribe}
Feature Bank	<awesome, movie, hollywood, celebrity, ...>
Binary	<awesome, movie, hollywood, celebrity, ...> : < 1, 1, 0, 0,... >
Count	<awesome, movie, hollywood, celebrity, ...> : < 1, 2, 0, 0,... >
TF-IDF	<awesome, movie, hollywood, celebrity, ...>: < 1.12, 1.47, 0, 0... >

Table 6. Feature extraction and selection results

Dataset (Lexical Feature)	N of feature	
	Feature Extraction	Feature Selection
Title	1908	142
Description	11186	970
Comments	38970	3688
Title + Description	11508	982
Title + Comments	39208	3689
Description + Comments	43641	4051
Title + Description + Comments	43768	4054

RESULT AND DISCUSSION

In order to produce the model for the classifier component, some experiments have been conducted. The goal of this experiment is to produce the best model in restricted content classification based on metadata and video comments. The experiments were conducted on Weka 3.7.9 [5]. Algorithm used for the model trainings are Naive Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF), which are provided in Weka.

Naive bayes works by calculating each attributes probability on each class. The classification of a new data by naive bayes done by finding the class with the highest probability based on those attributes [7].

SVM, on the other hand, works by projecting the feature vector into n-dimension space and creating a hyperplane which separate two classes with the highest margin based on the support vectors. New data can be classified by using the hyperplane that has been created by the SVM. To separate non-linear problems, SVM use the help of non-liner kernel such as RBF and Polynomial [7].

RF works by creating multi decision tree that is trained using different features. Each tree is constructed by giving each of them unique features that is produced by using bagging (bootstrap aggregating) method [8]. New data will be classified based on the majority results of each decision trees [7].

Table 7. F-Measure on training data using 10-folds cross validation

Feature Dataset (Lexical)	Weight	NB	SVM	RF
T	Binary	91.61%	91.46%	93.57%
T	Count	91.81%	89.04%	93.57%
T	TFIDF	91.61%	91.46%	93.55%
D	Binary	85.25%	91.37%	91.11%
D	Count	76.70%	71.71%	91.95%
D	TFIDF	85.25%	91.37%	91.10%
C	Binary	81.01%	89.39%	83.60%
C	Count	85.11%	84.10%	84.02%
C	TFIDF	81.01%	89.39%	83.84%
T + D	Binary	87.40%	**93.88%**	92.80%
T + D	Count	83.31%	75.32%	92.59%
T + D	TFIDF	87.40%	**93.88%**	92.46%
T + C	Binary	82.90%	91.69%	87.89%
T + C	Count	86.79%	84.87%	88.02%
T + C	TFIDF	82.90%	91.69%	87.93%
D + C	Binary	84.29%	92.26%	87.60%
D + C	Count	88.91%	87.58%	88.02%
D + C	TFIDF	84.29%	92.26%	87.49%
T + D + C	Binary	82.95%	91.93%	86.41%
T + D + C	Count	88.28%	87.33%	86.75%
T + D + C	TFIDF	82.95%	91.93%	85.69%

Each model will be evaluated using F-Measure as the measurement method [9]. The result of the experiments on training data can be seen on Table 7. T means title, D means description, and C means comments.

Table 7 shows the results of experiments conducted on training data using 10 folds cross validation [10]. Model with highest F-Measure of 93.88% was achieved using title and

description as the lexical feature source, SVM as the classification algorithm, with binary or tf-idf method as the feature weighting method. The testing result on table 8 showing some consistency with the same models achieving the highest F-Measure of 83.45%.

Table 8. F-Measure on testing data

Feature Dataset (Lexical)	Weight	NB	SVM	RF
T	Binary	79.90%	76.53%	79.81%
T	Count	71.40%	70.77%	79.77%
T	TFIDF	79.70%	76.53%	79.21%
D	Binary	80.34%	83.15%	79.76%
D	Count	82.64%	75.35%	80.34%
D	TFIDF	80.34%	83.15%	79.35%
C	Binary	75.10%	76.92%	76.97%
C	Count	75.18%	74.43%	77.03%
C	TFIDF	75.10%	76.92%	76.24%
T + D	Binary	80.11%	**83.45%**	80.18%
T + D	Count	82.18%	74.44%	82.29%
T + D	TFIDF	80.11%	**83.45%**	78.95%
T + C	Binary	76.39%	78.24%	80.07%
T + C	Count	78.76%	75.24%	79.67%
T + C	TFIDF	76.39%	78.24%	80.07%
D + C	Binary	76.08%	77.77%	79.80%
D + C	Count	78.69%	75.27%	81.07%
D + C	TFIDF	76.08%	77.77%	79.58%
T + D + C	Binary	74.47%	76.05%	74.94%
T + D + C	Count	76.04%	73.71%	77.51%
T + D + C	TFIDF	74.47%	76.05%	77.37%

The experiments results show that title and description hold the most valuable information in restricted content classification. This results can be explained by the nature of title and description which are written/produced by video uploader. The fact that video title and description is the key for video searching on youtube caused the uploader to write the title and description as relevant as possible with less typographical errors compared to the comments written by the user/watcher. The experiments also show that the video comments aren't as valuable as we expect it to be. The main factor that caused this is the fact that video comments section has no guideline which makes it often irrelevant and full of typographical errors. The facts that the number of comments on each videos varies a lot and the imbalanced amount of comments on restricted videos compared to non-restricted videos should be counted to make the classification model. More experiments related to those hypothesis can be found on Figure 3.

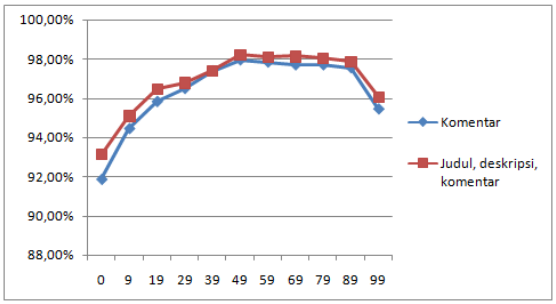

Figure 3. Relation between number of comments to *F-Measure*

Figure 3 shows that video with higher comments (49-89) have higher information quality to video with less comments. The reduction of F-Measure as the number of comments go higher than 49 should be caused by the fact that restricted videos in general has less comment compared to non-restricted videos thus makes the dataset more imbalanced as the minimum number of comments increased.

Table 9 Correctly classified video sample

Video Data	Content
URL	http://www.youtube.com/watch?v=SjHJuALwvM0
Title	Male To Female Full Transformation
Description	Subscribe Please http://www.youtube.com/subscription_center?add_user=RBuTubeV
Rating	Safe (Non-Restricted)
Comments	- i think this kid is pretty talented..,make a good make-up artist one day.. - What's a trabsgenders sexuality - How did u do the boobies - im wonding wat the point u go wit dress as a girl? u go to clubes and get fuck im the booty hole or you just suck dick - this is disgusting as SHIT...!! - My lord. your a fucking faggot - ...etc

Table 9 shows the sample of the correctly calssified video using the best model achieved from experiments (T+D). On the other hand, using video comments as another feature source (C, T+C, T+D+C, etc) would cause this

video to be wrongly classified since the comments section contain many features that fit restricted category (offensive and sexual words).

Table 10 Incorrectly classified video sample

Video Data	Content
URL	https://www.youtube.com/watch?v=5S8tMi7woc4
Judul	TAZ DYESS Ft. NOONYY - Sage The Gemini - Red Nose "Yike Dance"(Grind Video Version)
Description	Follow Noonyy On InstaGram : @Noonyyyy or Click https://Instagram.com/Noonyyyy Want TAZ DYESS To Come to yo City (SERIOUS BOOKINGS ONLY!!!!!!) Contact my Manager shay at 706 619 6085 or email shaysweetpeach27@Gmail.com MY Personal Contacts Twitter Etc..(BELOW) INSTAGRAM: @Only1TazDyess or Click https://Instagram.com/Only1TazDyess ASK.FM: http://ask.fm/TheRealTazDyess TWITTER: @Only1TazDyess Or Click https://twitter.com/Only1TazDyess ...etc
Rating	Restricted
Comments	- nice and sexy - wat da fuck was they doin they was having sex thats wat they should called it a sex video. - Dont hate i q ishbi was here and you girls know you do to because he is fine - It ain't called hatin if it's ur opinion - The girl cakeyy but its obvious sex was involved after - I am 10 and I know that kids is not suppose to see this video but what is you doing that's not the red nose - ...etc

Table 10 shows the sample of incorrectly classified video using the best model from experiments (T+D). The model fails to classify this video sample because of the lack of information found on the title and description. Judging by the title and description of the video, it looks like just another music video while the content of the actual video depicted some action that is categorized as restricted content by Youtube.

Using the best model achieved from the experiment, we then proceed to built a working prototype of the classification system. The prototype was built as a web apllication to make it easily accessible by the user. Screenshot of the working prototype can be seen on Figure 4.

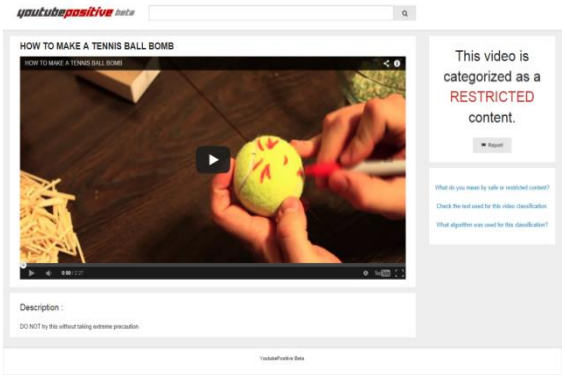

Figure 4. Screenshot of the classification system prototype

CONCLUSION

In this research, we have successfully developed a prototype for automatic restricted content classification using the best model acquired from experiments. The best model has 83.45% F-Measure achieved by using video title and description as the source of lexical feature with SVM as the classification algorithm and binary as the feature weighting method. The prototype is built as a web application with three main components which are data crawler, classifier, and user interface.

REFERENCES

[1] Youtube. Statistics. URL: https://www.youtube.com/yt/press/statistics.html, accessed on December 2014

[2] Youtube. Age-restricted content. URL: https://support.google.com/youtube/answer/2802167?hl=en, accessed on December 2014.

[3] K. Filippova, K. B.Hall, "Improved Video Categorization from Text Metadata and User Comments," *Special Interest Group in Information Retrieval* (SIGIR), vol. 5, no.7, 2011.

[4] R. Du, R. Safavi-Naini, and W. Susilo, "Web filtering using text classification," *in Proceedings of The 11th IEEE International Conference on Networks*, pp. 325-330, 2003.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[6] A. Moschitti, R.Basili, "Complex Linguistic Features for Text Classification: A Comprehensive Study," *Springer Verlag*, vol. 5, no.10, 2004.

[7] T. Mitchell, Machine Learning, New York : McGraw-Hill, 1997

[8] L. Breiman, "Bagging Predictors. Machine Learning," vol. 24, no.2, pp. 123-140, 1996.

[9] D. M. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness,Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.

[10] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *in Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, vol. 2, no.12, pp. 1137–1143, 1995.