# Identifying Suitable YouTube Videos for Children

Carsten Eickhoff [1], Arjen P. de Vries [2]

[1]TU Delft, Delft, Netherlands; [2]Centrum Wiskunde & Informatica, Amsterdam, Netherlands

E-mail: [1] c.eickhoff@tudelft.nl, [2] arjen@acm.org

*Abstract:* **Content-sharing platforms such as YouTube or MyVideo are experiencing huge user numbers that are still rising very quickly. Among the users there is a steadily growing share of children. In spite of this tendency the content of many popular videos is not suitable for children and should therefore not be shown to them. In this work we present an automatic method for determining a shared video's suitability for children based on non-audio-visual data. We evaluate its performance on a corpus of web videos that was annotated by domain experts. We finally show how community expertise in the form of user comments and ratings can yield better prediction results than directly video-related information.**

**Keywords:** Classification, Content Filtering, Video, Children, Content Mining

## 1 INTRODUCTION

For several years now content-sharing platforms on the Internet represent one of the pillars of multimedia society. The connection of sharing content and offering social networking features, as for example the ability to comment on content, appeals to a broad range of users. Video sharing platforms are particularly popular and have experienced great growth rates in both, the amount of shared content as well as the number of viewers. A recent survey attributed the video-sharing platform YouTube as being solely responsible for approximately 10% of the global Internet traffic [1].

At the same time, pedagogues and social scientists notice that the age of first exposure to computers in general and the Internet in particular has been rapidly decreasing with the overall amount of Internet and media consumption of children rising [2]. An additional factor of attraction towards video sharing for very young Internet users lies in the nature of the content. While textual resources often require considerable literacy skills, videos also appeal to children who can not yet read very well. The potential danger lies in the unmoderated consumption of videos. Parents or teachers who guide children's information seeking naturally filter which content to show to their wards and which to avoid. A British study on the media consumption of UK children however finds, that in practice children nowadays often have unmoderated access to computers and the Internet [3]. They found that as much as 40% of children aged between 5 and 15 years regularly access the Internet without adult supervision or guidance.

Video-sharing communities that specifically target young audiences appear to be a promising alternative. Manually selected collections of videos as for example offered on Totlol [4] provide high quality content for children and parents. They do however impose a high work load of manual editing and selecting on the community. Hand-picked collections typically feature a comparably low coverage rate and low agility as it takes time for the community to explore the available content.

Since permanent assistance of a prudent adult can in reality not always be ensured and hand-selected video collections require high maintenance efforts, an automatic method for identifying suitable video resources is desirable. In this work we propose an approach that makes use of the existing user annotations on video sharing platforms. In this way we are able to transform the implicit human judgements into suitability predictions.

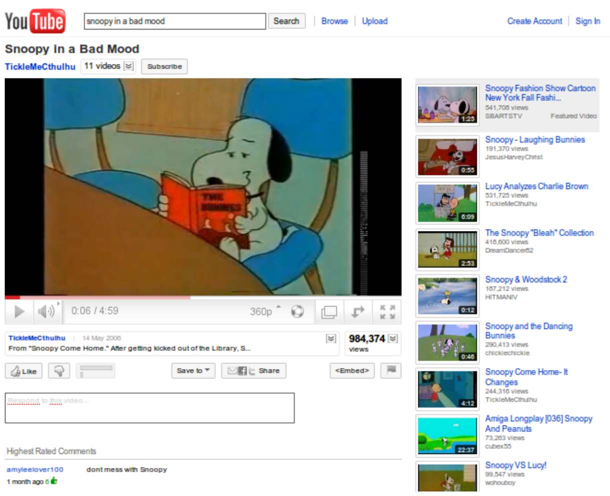


**Figure 1: YouTube video structure**

The contributions of this work are threefold: (1) We discuss and introduce the previously unstudied task of automatically determining age suitability of shared web videos. (2) We describe a range of available features and motivate their use for the task at hand. (3) We conduct a range of experiments showing that community information can yield stronger clues about the nature of

**Corresponding author:** Carsten Eickhoff, TU Delft, Mekelweg 4, 2628CD Delft, Netherlands, +31 (0)15 27 81845, c.eickhoff@tudelft.nl

content than pre-made content information such as tags or full-text descriptions.

The remainder of this work is structured as follows: Section 2 gives an overview of previous work in related fields. Section 3 introduces and motivates a range of features that are assumed to be strong indicators of video suitability. Section 4 describes our experimental corpus of videos and explores the domain of shared web videos for children. Section 5 will finally draw a conclusion and discuss potential future routes to pursue.

# 2 RELATED WORK

While there has been no previous work dedicated to determining the suitability of web videos for children a range of relevant research on related topics has been conducted.

In recent years, much research effort has been invested into automatic video classification. Traditional video classification approaches at first commonly employed audio-visual features [9] often using Hidden Markov Models [10, 11]. There have however been promising advances into using more sophisticated machine learning techniques such as Support Vector Machines [13]. The growing amount of shared and tagged video content available has given rise to approaches making use of textual features [12]. The fundamental difference to our method however lies in the objective. While topic information appears to be well-contained in tags and headlines, age appropriateness is widely independent of the video's subject.

Another closely related topic is the predictive and indicative potential of user comments on content-sharing platforms. In 2008, Lange [5] found that user comments on YouTube often more clearly express the relationships between users than the platform's explicit friendship function. Yee et al. [6] investigated whether user comments on a content-sharing site can be used to improve search performance. They showed that including user comments into the search index can yield significant accuracy gains. Siersdorfer et al. [7] asked the more general question of the overall usefulness of YouTube user comments. In several experiments they were able to build models to identify polarising comments as well as predicting comment ratings. De Choudhury et al. [8] finally tried to identify characteristics of interesting conversations by analysing discussions in YouTube comments. These various successful exploitations of information contained in comments encourage our approach of determining suitability of video content based on user comments.

# 3 FEATURE EXTRACTION

As a first step towards identifying suitable video content for children this section will introduce a range of potential features that may convey suitability information.

## 3.1 The structure of YouTube content

We will start with a brief inspection of the information offered on a typical YouTube page. An example is shown in Figure 1. The pieces of information on the page can be grouped into 4 distinct categories:

**Video information** This category contains all information concerning the actual video content. Examples from this category are the video title, tags, full text descriptions, its genre or the play time. Intuitively one would assume this category to contain the strongest indicators of suitability as it is directly dedicated to the video content. We will however show in the course of this work, that other sources of information can yield comparably strong indications.

**Author information** The second source of information available on the page is related to the video author. While on the actual video page we can only find the author's user name, looking at the related user profile offers many interesting clues.

**Meta Information** This category is dedicated to automatically generated meta information on the video. Specimen from this group are the video's publication date or the number of times it was viewed.

**Community-created information** The fourth category is concerned with all information that was created by the user community. Examples are user ratings, video responses and comments. In this work, we will put strong emphasis on the community aspect and its use for determining video suitability.

## 3.2 Capturing Suitability

After having introduced the range of information available on YouTube pages we will now discuss their applicability for predicting suitability of video content.

### Video Information

The main textual resources in this category are the video's tags and its description. The title and the category turn out to be already contained term by term in the list of tags. We propose using uni-gram tag-based language models to represent topical similarities and tag co-occurrences of child/non-child videos. Due to the relative brevity of the video descriptions, we expect only limited growth in complexity when using higher order models at this point.

Another feature of interest is the video's play time. Topical classification was hardly able to exploit this aspect. Children however show a significantly lower attention span than adults [16]. We will investigate whether this is reflected in the length of children's videos compared to general video clips.

### Author Information

User profiles offer a wide range of possibly interesting information. The user's age may be relevant towards the suitability decision of her or his content. It is likely that user age forms a valid prior on suitability as children's videos will in the majority of cases be watched by actual children or by parents which fall into empirically distinguishable age groups.

Furthermore the profile text (either free text or information on favourite books, films, hobbies, etc.) is expected to provide valuable information on the user's

**Corresponding author:** Carsten Eickhoff, TU Delft, Mekelweg 4, 2628CD Delft, Netherlands, +31 (0)15 27 81845, c.eickhoff@tudelft.nl

background and interests. We will use a language modelling approach to reflect these specifics for authors of children's and general content. It is however important to consider the potentially sparse nature of these texts. The user is not bound to provide and maintain a detailed profile of his likes and dislikes.

### Meta Information

An initial exploration of a number of videos showed no significant correlation between this category and the suitability for children. For reasons of completeness and comparability we still included this category in the further experiments of this work.

### Community-created Information

#### Topicality

This category, finally, is the main object of interest for our work. YouTube comments are typically short user messages, related to the published video content (although we also encountered extensive paragraphs of text). Users additionally have the possibility to rate comments. The highest-rated comments are then shown in a priority position above the main discussion. As a first step towards using comments for suitability prediction we will build comment language models for children's videos as opposed to general videos. These models are expected to give further evidence of the video's topic and related subjects.

#### Controversy

One of the strongest motors of commenting is controversy. People feel more far more inclined to contribute to a discussion whose general position they disagree with. This results in controversial topics being discussed more vividly with comments pouring in every few seconds [5]. Considering the nature of children's videos we expect to see far fewer heated debates on an episode of “Hello Kitty” than there might be on a news story that deals with the recent changes to the US healthcare system. As a consequence we will consider the total number of comments but also the median time between comments as features.

We will furthermore capture this notion by applying sentiment analysis to the video comments. Children's content is in general expected to cause positive affection rather than negative sentiments. The typical behaviour of antagonism that is often observed in on-line discussions [14] is expected to be less frequent for children's content. Our sentiment analysis is based on the sentiment corpus by Pang and Lee [15]. The likelihood of a given comment being well-intentioned is expressed as the average of its constituent term's likelihoods. The likelihood of each single term t being positive is finally defined as the number of times t was observed in a positive comment in relation to the total occurrences of t.

$$P(positive|c) = \frac{1}{|c|}\sum_{i=0}^{|c|} p(positive|t_i)$$

$$p(positive|t) = \frac{count_{positive}(t)}{count_{total}(t)}$$

The analogous score for the negative case is computed as well and both are reported as features.

A complete overview of all features and their category affiliations is given in table 1.

## 4 EXPERIMENTS

### 4.1 Data Set

Our research corpus consists of 12,673 YouTube videos which were collected in January 2010. The crawling process was started from 20 distinct queries that were deemed typical specimen of either children's queries (e.g. “Hello Kitty” or “Sesame Street”) or non-child ones (e.g. “Iraq war” or “Klitschko boxing”). For each of the seed queries the top 3 videos were visited and crawled. The crawling step collected the relevant meta data to extract the features described in the previous section. Starting from the initial set at each point we queued the top 5 related videos for successive crawling.

**Table 1: YouTube features**

| Video Information | Author Information | Meta Information | Community Information |
|---|---|---|---|
| Play time | Age | Publication date | Average rating |
| Tag LM (1-gram) | Profile text LM (3-gram) | # of views | Uncapped # of comments |
| Description LM (3-gram) | # of subscribers | | Comment LM (3-gram) |
| Presence of tag “kid” | # of views of published videos | | Sentiment score positive |
| Presence of tag “child” | | | Sentiment score negative |
| | | | # favourites |
| | | | Median inter-comment interval |

While the average video in our collection had a total of 360 comments, there are outliers which solely feature as many as 281,571 comments for famous pieces of popular culture. In order to reduce the crawling and processing time of such videos to reasonable dimensions we capped the number of comments fetched at $\delta_{threshold}$ = 4950 comments. At that point more than 96.8% of the videos do not have additional comments. Without effecting most of the videos the computational load could thus be significantly reduced.

Out of the whole collection an initial sample of 1000 videos (~50% suitable for children and ~50% for adult audiences) have been rated concerning their child-friendliness by a domain expert with a background in childcare and education.

**Corresponding author:** Carsten Eickhoff, TU Delft, Mekelweg 4, 2628CD Delft, Netherlands, +31 (0)15 27 81845, c.eickhoff@tudelft.nl

## 4.2 Domain Exploration

In this section, we will explore the domain of child-friendly shared web videos in the course of a number of experiments. Our exploration will be guided by the following three research questions: (1) Can we automatically identify child-friendly videos using exclusively non-audio-visual information? (2) Can community expertise outperform directly video-related information at predicting child-friendliness? (3) Can video play time indicate child-friendliness of shared web videos?

To begin our inspection of the domain we split the corpus into stratified training (90%) and test (10%) sets, extracted the previously described features and trained a range of state of the art machine learning classifiers. Table 2 shows a performance comparison of the various approaches on the previously unseen test set. Performance will be captured in terms of precision and recall as well as their combination in the F0.5-measure. We decided for the precision-biased F score to reflect the nature of our task. Showing as few as possible unsuitable videos to children (high precision) is clearly more important than retrieving all suitable videos (high recall). The area under the ROC curve is additionally reported to give a notion of classification confidence.

**Table 2: Classification performance comparison**

| Classifier | P | R | $F_{0.5}$ | ROC |
|---|---|---|---|---|
| SVM | 0.85 | 0.67 | 0.81 | 0.85 |
| Random Forest | 0.77 | 0.86 | 0.79 | 0.87 |
| Decision Table | 0.75 | 0.83 | 0.76 | 0.79 |
| Logistic Regression | 0.72 | 0.63 | 0.7 | 0.74 |
| Decision Tree | 0.74 | 0.79 | 0.75 | 0.82 |
| Ada Boost | 0.72 | 0.89 | 0.75 | 0.79 |
| MLP | 0.78 | 0.6 | 0.74 | 0.77 |
| Naïve Bayes | 0.72 | 0.63 | 0.7 | 0.72 |

We can observe that already straight forward Naïve Bayesian approaches yielded convincing performance. The overall best-performing model was an SVM classifier (We employed an SVM using a Pearson VII universal kernel function as suggested by Qifu et al. [17]. The following parameter settings were used: $\omega = 1$, $\sigma = 2$, $\varepsilon = 10^{-12}$ and c = 1.) Although we did not invest time into further feature engineering and parameter tuning at this point of our research, the results look promising.

**Table 3: Feature category performance comparison**

| Category | P | R | $F_{0.5}$ | ROC |
|---|---|---|---|---|
| Video-related Information | 0.61 | 0.79 | 0.64 | 0.63 |
| Author Information | 0.65 | 0.79 | 0.67 | 0.69 |
| Meta Information | 0.66 | 0.54 | 0.63 | 0.61 |
| **Community Information** | **0.7** | **0.72** | **0.7** | **0.73** |

We thus see our first research hypothesis confirmed; Suitability of shared videos can be reliably estimated using exclusively non-audio-visual features. In order to gain a deeper understanding of how the crucial information is reflected in the data, we will further analyse the individual prediction performance per feature category and per single feature. For this purpose we again used the previously described SVM which was now however trained on one feature category at a time. Table 3 shows a ranking by feature category performance. As expected, meta information turned out to be the weakest overall category. Video-related information performed worse than both author information and community-generated information. Returning to our second research question, we note that community information which was the strongest feature category represents a more powerful predictor of suitability for children than directly video-related features. This finding is statistically significant at $\alpha < 0.05$-level (determined using Wilcoxon Signed-Rank test).

We conducted the analogous experiment training the classifiers on just a single feature at a time. The results are shown in table 4.

**Table 4: Single feature performance comparison**

| Feature | P | R | $F_{0.5}$ | ROC |
|---|---|---|---|---|
| # of views | 0.75 | 0.54 | 0.7 | 0.72 |
| Average rating | 0.66 | 0.85 | 0.69 | 0.62 |
| # favourites | 0.72 | 0.54 | 0.68 | 0.7 |
| # of views of published videos | 0.72 | 0.54 | 0.68 | 0.67 |
| Median inter-comment interval | 0.71 | 0.55 | 0.67 | 0.7 |
| Author age | 0.64 | 0.82 | 0.67 | 0.64 |
| Tag LM (1-gram) | 0.59 | 0.92 | 0.64 | 0.55 |
| Profile text LM (3-gram) | 0.6 | 0.8 | 0.63 | 0.65 |
| Comment LM (3-gram) | 0.58 | 0.87 | 0.62 | 0.56 |
| Sentiment score positive | 0.57 | 0.91 | 0.62 | 0.52 |
| Sentiment score negative | 0.57 | 0.88 | 0.61 | 0.54 |
| Description LM (3-gram) | 0.56 | 0.91 | 0.61 | 0.53 |
| # of subscribers | 0.55 | 0.99 | 0.6 | 0.56 |
| Presence of tag “kid” | 0.55 | 1 | 0.6 | 0.54 |
| Uncapped # of comments | 0.55 | 0.98 | 0.6 | 0.5 |
| Presence of tag “child” | 0.55 | 1 | 0.6 | 0.5 |
| Publication date | 0.52 | 1 | 0.58 | 0.53 |
| Video Play time | 0.54 | 0.48 | 0.53 | 0.51 |

Closer examination of the results shows a surprising tendency. The strongest single feature turned out to be the number of times the video was watched. This finding is most likely due to the fact that the majority of extremely popular videos on YouTube are about current pop culture. The number of people who are actually interested in watching children's videos will be limited in comparison to the fan community of a famous musician or music style. While this finding represents an interesting foundation for further research it is of course not prudent to deduce that every video that is not largely popular in

**Corresponding author:** Carsten Eickhoff, TU Delft, Mekelweg 4, 2628CD Delft, Netherlands, +31 (0)15 27 81845, c.eickhoff@tudelft.nl

terms of number of views could be shown to children without concern.

Reconsidering our third research question we have to note that (at least at the moment) the mere duration of a YouTube video does not give valid clues to its suitability for children. Despite the findings of Anderson et al. [16] who were able to measure significant differences in the attention spans of children of different age groups, the hypothesis of children's videos being shorter did not hold true for our application. We assume that this is largely due to the fact that video clips on YouTube are short by definition. Video sequences of no more than 10 minutes are apparently easily understandable for children so that no significant differences could be measured. In the future, this tendency might however be subject to changes if YouTube should decide to lift or loosen the 10 minute duration limit. In that case the differences in video length for children and adults should be re-evaluated.

# 5 CONCLUSION

This work represents an initial exploration of automatic suitability prediction of shared web videos. We used the video sharing platform YouTube as an example and described the various sources of information available on their pages. We conducted a feasibility study of automatically classifying YouTube videos into those suitable for children and those which are unsuitable. We found that even our initial approaches yielded reliable results.

In our second experiment, we could confirm the important role of community expertise for determining suitability of videos. Models based on community-generated information such as user comments, ratings and favourite declarations were able to perform significantly better than those exclusively built on content-related information such as tags and genres. This tendency is very promising for future research as we expect related fields to benefit from the same observation. In the area of content-sharing the (often tedious) task of tagging might become less and less important if we gain a better understanding of how to make use of the information that the community generates naturally. At that stage one would only have to resort to content analysis for very new content that has not yet been discussed by the community. Such an unobtrusive way of relating concepts as for example genre or suitability to content would be highly desirable.

This work describes a piece of research in progress and a great number of promising future directions are yet to be pursued. The first step towards a broader understanding of the domain will be to generalize our findings from YouTube to the abstract case of general community-shared video content. Recently YouTube introduced a new way of expressing affection towards a given video or comment in the form of “like” and “dislike” buttons. This quick and popular form of feedback was not present when we collected the data set for our current research. It should however be considered for future work as it might give the user community another valuable tool to express their expertise. Finally we regard the networking aspect of content-sharing platforms to be highly beneficial. Following related videos, video answers and additional uploads by the author or people who liked the content may give strong indications of the suitability of the actual video. Similar approaches have been shown to perform well for the topical classification of web sites [18] and should therefore be evaluated in this domain as well.

## Acknowledgement

This research is part of the PuppyIR project [19]. It is funded by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507.

## References

[1] X. Cheng, C. Dale and J. Liu. Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study. *ArXiv e-prints*, 2007

[2] E.A. Wartella, E.A. Vandewater and V.J. Rideout. Introduction: Electronic media use in the lives of infants, toddlers, and preschoolers. In *American Behavioral Scientist*, 48(5):501, 2005

[3] OfCom. UK children's media literacy: Research document. http://www.ofcom.org.uk/advice /media/literacy/medlitpub/ medlitpubrss/ukchildrensml/ukchildrensml1.pdf, March 2010

[4] R.Ilan. Totlol – Daddy, Mommy, more video! http://www.totlol.com/, 2010

[5] P.G.Lange. Publicly Private and Privately Public: Social Networking on YouTube. In *Journal of Computer-Mediated Communication* (13):361-380, 2008

[6] W.G.Yee, A. Yates, S. Liu, and O. Frieder. Are Web User Comments Useful for Search? In *7th Workshop on Large-Scale Distributed Systems for Information Retrieval*. Boston, USA, 2009

[7] S.Siersdorfer, S. Chelaru and W. Nejdl. How Useful are Your Comments? - Analyzing and Predicting YouTube Comments and Comment Ratings. In *WWW 2010*, Raleigh, USA

[8] M. de Choudhury, H. Sundaram, A. John and D.D. Seligmann. What Makes Conversations Interesting? Themes, Participants and Consequences of Conversations in Online Social Media. In *WWW 2009*, Madrid, Spain

[9] Y.Wang, Z.Liu, and J.C.Huang. Multimedia Content Analysis Using Both Audio and Visual Clues. In *IEEE Signal Processing Magazine 17(6)*, 2000.

[10] Z.Liu, J.Huang and Y.Wang. Classification of TV Programs Based on Audio Information Using Hidden Markov Model. In *IEEE Workshop on Multimedia Signal Processing,* 1998.

[11] C.Lu, M.S.Drew and J.Au. Classification of Summarized Videos using Hidden Markov Models on Compressed Chromaticity Signatures. In *ACM Multimedia, Ottawa, Canada*, 2001

[12] W.Lin and A.Hauptmann. News video classification using SVM-based multimodal classifiers and combination strategies. In *ACM Multimedia, Juan Les Pins, France*, 2002

[13] V.Suresh, C.K.Mohan, R.K.Swamy and B. Yegnanarayana. Content-Based Video Classification using Support Vector Machines. In *Neural Information Processing, Calcutta, India,* 2004

[14] P.G.Lange. Commenting on comments: Investigating responses to antagonism on YouTube. In *Annual Conference of the Society for Applied Anthropology, Tampa, USA*, 2007

[15] B.Pang and L.Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the ACL, Barcelona, Spain*, 2004

[16] D.R.Anderson and S.R.Levin. Young Children's Attention to “Sesame Street”. In *Child Development 47 (3)*, 1976

[17] Z.Qifu, H.Haifeng, Z.Youzheng and S.Guodong. Support Vector Machine Based on Universal Kernel Function and its Application in Quantitative Structure – Toxicity Relationship Model. In *International Forum on Information Technology and Application, Chengdu, China,* 2009

[18] X.Qi and B.D.Davison. Web page classification: Features and algorithms. In *ACM CSUR II*, 2009

[19] PuppyIR: An Open Source Environment to Construct Information Services for Children. http://www.puppyir.eu.

**Corresponding author:** Carsten Eickhoff, TU Delft, Mekelweg 4, 2628CD Delft, Netherlands, +31 (0)15 27 81845, c.eickhoff@tudelft.nl