Disinformation in Multimedia Annotation: Misleading Metadata Detection on YouTube

Payal Bajaj
Adobe Research Labs
Bengaluru, India
pabajaj@adobe.com

Mridul Kavidayal
Adobe Systems India
Noida, India
kavidaya@adobe.com

Priyanshu Srivastava
Adobe Systems India
Noida, India
psrivast@adobe.com

Md Nadeem Akhtar
Adobe Systems India
Noida, India
mdakhtar@adobe.com

Ponnurangam Kumaraguru
IIIT Delhi
Delhi, India
pk@iiitd.ac.in

ABSTRACT

Popularity of online videos is increasing at a rapid rate. Not only the users can access these videos online, but they can also upload video content on platforms like $YouTube$ and $Myspace$. These videos are indexed by user generated multimedia annotation, also known as metadata, which is usually rich contextual information added by users about the content of the videos to facilitate access to their videos. Metadata plays a crucial role in techniques for video search and retrieval. However, this freedom of choosing annotation causes some uploaders to provide additional tags which are not even related to the content of the videos. Therefore, it is essential to verify the relevance of user-generated tags with the content of the video. Given the sheer volume of video content uploaded everyday, manual tag validation can be a highly labor intensive task. In this paper, we propose a method to automatically analyze user generated tags against video content to identify relevance of these tags and to detect irrelevant and misleading metadata for online videos.

Our contributions are three-fold: First, we study nature of user-assigned tags and characterize them in two categories-$generic$ and $specific$ tags. Second, we propose a novel hierarchical graph based approach to identify tags which are relevant to content of the video. Third, we present a way to use user-generated comments for multimedia annotation verification. We demonstrate results of our method and evaluation on 300 $YouTube$ videos for three different categories. The results show that we are able to identify relevant tags with average recall of 0.813 and average precision of 0.97.

CCS Concepts

•Information systems → Multimedia information systems;

iV&L-MM'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2983563.2983569

Keywords

Multimedia Annotation; Disinformation; Video Tags

1. INTRODUCTION

In the past decade, the Web has become widely accessible and the most crucial source and trusted platform to share as well as retrieve user generated content, such as reviews and videos. Given the increasing reliance of users on Web, it is essential to keep a check on the authenticity and quality of this easily accessible information. This problem has attracted a lot of attention in the past because the quality of information directly affects user experience and website credibility.

There has been a lot of focus of verifying textual content shared by users, such as credibility of social media posts [9] and misinformation on Wikipedia [27]. However, there has not been much focus on verifying textual information in context of videos. This paper specifically focuses on verification of metadata for tagged images or annotated videos. The online video industry has seen a massive growth in recent years. Proliferation of platforms like $YouTube$ and $Myspace$ has increased users' interaction with video content online. The users can not only access and share video content online, but also upload a variety of videos themselves. The uploaded videos can include all kinds of video content spanning various genre. Therefore, generally all such platforms allow the uploaders to assist their content with tags, metadata, title and description to facilitate better video search results for browsing and retrieval.

Given the sheer volume of video content generated by users[36], it is not hard to realize that the user generated metadata and tags play a crucial role in video search and retrieval algorithms. So, the textual content provided by the users in form of tags directly impacts the number of impressions and views of a particular video. Whereas there can be innocent attempts to add related tags to make the video more accessible, this opens up a possibility for the users to deliberately include tags that are not related to video content just to boost video popularity and acquire more eyeball impressions. For instance, users can add tags corresponding to trending topics and alter the thumbnail appropriately so that their video appears whenever viewers query for popular videos with trending topics, whereas the actual content

of the video might not be even related to those trending tags. Thus, on one side, the flexibility of providing video content with metadata assists with video content but on the other side, might misguide viewers by misleading metadata. Therefore, validating the tags provided by users on uploaded videos is an important problem and demands research focus.

Tag verification is not a straightforward task. This is primarily because user generated metadata can span a large variety of concepts. On one side, there can be specific concepts such as *basketball*, *cats*, *dogs*, and proper nouns for movie stars, sports players, etc., which can be detected by predicting the objects and concepts present in the videos using image processing techniques. But on the other side, there can be tags as high level and generic as *movie*, *game*, and *sports* which might not have standalone image detection mechanisms. For instance, a $YouTube$ video might mention the tags *Game*, *Sports*, *Stephen Curry* and *Basketball*. Though we can verify the last two tags as mentioned before, but we need a mechanism to verify the first two generic tags also to solve the problem.

In addition to handling generic and abstract keywords, tag ambiguity and associations introduce additional complexity in tag verification task. The tags like *ocean*, *river* and *sea* are associated with each other, but the image classification techniques might detect only one of these and mark the other two tags as irrelevant, which might not be the case. We need a mechanism to handle such tag associations - polysemy, synonyms, hypernyms and hyponyms for accurate validation of tags. In fact, the problem can't necessarily be handled by creating a hard-coded list of similar meaning words together because this might associate words even out of context. For instance, *bat* and *cricket* might or might not be related depending on the content of the video, whereas *usa* and *america* are always related. Therefore, it is important to address the meaning of these keywords in context of the video.

Validation of tags is a highly manual process now. For instance, $YouTube$ offers its users to report cases of misleading metadata, following which the video and its tags are examined manually and the video is removed if infact the tags were irrelevant. This process is a highly manpower intensive as it requires content maintainers to go through the entire video before deciding which tags are relevant. We propose a method to automatically validate the relevance of tags annotated by a user for a particular video with respect to its content. This will facilitate verification of the metadata users associated with the video and identify fraudulent and irrelevant tags without manual intervention.In this paper, our work focuses on detecting misleading or irrelevant tags associated by the users with online videos. We provide a systematic study of detecting misleading tags for online videos. To summarize, the contributions of this paper are - (1) To the best of our knowledge, we are the first one to study the verification of user-generated metadata for online videos as a problem of identifying relevance of tags with respect to content of the video. (2) We consider that users associate specific as well as generic tags with the videos. Hence, we provide a mechanism to address all variations by establishing tag relationship through corpus co-occurrence modeling. (3) Based on the content of the video and tag relationships, we identify which of the tags are actually relevant with respect to the content of the video. (4) Finally, we present how user generated comments can be used for tag verification process.

The rest of the paper is organized as follows. Section 2 describes the prior literature related to tag processing for videos. We formally define the problem statement in section 3 and discuss the proposed approach in section 4. Section 5 summarizes the dataset and experiments which is followed by results and evaluation in section 6. We discuss the applications of the proposed approach in section 7 and we conclude the paper in section 8.

2. RELATED WORK

In this section, we describe the prior literature on tag processing for online video content. With the increasing popularity of video platforms online, tag processing has gained increasing research interests to make video search and retrieval easier. We broadly classify the research in this area into two sub-categories below. We also summarize the research on identification of irrelevant metadata briefly.

2.1 Misinformation Detection

With the increasing popularity of the Web, assessing credibility of information on the web is becoming increasingly important and gathered a lot of research focus ([6], [15]). Most work in this domain revolves around detection and mitigation of misinformation on Wikipedia ([27]) and social networks [9]. Adler et al. develop a content-based reputation system for authors based on their edits by observing if their edits are rolled back or undone in short order in [10]. Tambuscio et al. present a simple modeling framework to study the diffusion of hoaxes in social networks and quantitatively gauge the minimal reaction necessary to eradicate them [11].

We also briefly discuss the issue of misinformation by mentioning malicious behaviour on YouTube platform. This is not directly related to our problem, but relates to the broader issue of false promotion of videos in YouTube. The increasing popularity of user generated multimedia content as a result of popular video hosting sites like $YouTube$ and its subsequent monetization drives uploaders to fraudulently promote their videos. Benevenuto et. al. [2] present a method to detect spammers who post an unrelated video as response to a popular one, aiming at increasing the likelihood of the response being viewed by a larger number of users, and content promoters who try to gain visibility to a specific video by posting a large number of (potentially unrelated) responses to boost the rank of the responded video, making it appear in the top lists maintained by the system. Bulakh et. al address a similar problem of identification of fraudulent promotion of videos in [3] by training supervised machine learning classifiers that can successfully differentiate fraudulent videos and profiles from legitimate ones.

There has been work to resolve uncontrolled, and ambiguous tagging by learning tag relevance by accumulating votes from visually similar neighbors ([17], [18]). However, there has been no work in identifying misinformation in the form of tags assigned by the uploaders in video sharing platforms like YouTube. To the best of our knowledge, we are the first to present a systematic study of misleading tags and their detection in YouTube videos.

2.2 Tag Processing

Tag processing has received it's fair share on interest from

the research community in the form of tag annotation ([20], [21], [22], [24], [25]), localization ([4]), recommendation ([8], [7]) and expansion ([13], [14], [8]).

However, all these techniques are limited to conceptual tags which can be detected using image processing such as objects, actions and scene models and can't be extended to broad category of abstract and generic classes such as *movie* and *game* for tag verification. This is primarily because of two reasons - (1) Tag associations introduce many complexities: Automatic image annotation technique might identify *ocean* in the image, and mark *sea* incorrect, and though these tags are related but might not be recognized due to polysemy. Therefore, we need to use semantics similarity between tags in order to verify them, and can't just rely on image processing techniques. (2) Inability to handle generic and abstract tags: The annotation schemes usually generate tags for the objects that can be identified in the images. If we take this output as ground-truth and mark all the other tags as irrelevant, the algorithm loses out on handling generic tags like *movie*, *sports*, *game* and *music*, even if it identifies conceptual tags like *Harry Potter*, *basketball*, *cricket match* and *piano* respectively.

Also, the previous work [16] has primarily relied upon existing taxonomies like WordNet[1] to extend identified concepts to more generic terms. It contain logical statements such as *songwriters* are *musicians*, that *musicians* are *humans* and that *they* cannot be any *other species*, or that *Canada* is part of *North America* and belongs to the *British Commonwealth*. However, this does not incorporate newer information such as *Bob Dylan* and *Leonard Cohen* are *song writers*, *Cohen* is born in *Montreal*, that *Montreal* is a *Canadian city*, or that both *Dylan* and *Cohen* have won the *Grammy Award*. This has been incorporated in present knowledge base [35] but information about the enormous class of songs, movies, celebrities and latest events like a match played yesterday is missing from these too. Given that user generated content is so diverse that their videos can be about any topic and can have a wide multitude of tags, knowledge bases are not adequate for the purpose of identifying all kinds of tags. Therefore they can't be used in a generalized way for tag verification. We use a web-based corpus-modeling technique to address these drawbacks.

There has been some work to address the problem of irrelevant metadata in images. Xirong Li et al. use visual similarity of images to determine tag relevance for social image retrieval in [19]. They propose a neighbour voting scheme based on the assumption that if different people use the same tags to label visually similar images, the tags are likely to be relevant to the visual content it is describing. However, this completely relies on user tags and similarity of user provided images and not on actual image content.

However, to the best of our knowledge, there has been no work on automatically identifying the relevance of user-generated tags with respect to the content of the video. Also, there is no mechanism to automatically detect irrelevant tags and misleading metadata for online videos.

3. PROBLEM STATEMENT

In this section, we formally define the problem statement we are solving. Let the set of videos be denoted by V and the number of videos be n. For every video $v \in V$, users annotate it with a set of tags T_v and a title H_v. Let $n_v = |T_v|$. Our problem is to divide the tags in T_v into two mutually exclusive and exhaustive subsets $\forall\ v \in V$:

$$T_v = \{V_v \cup N_v\} \tag{1}$$

Here, V_v is the set of verified tags - that which could be identified as related to the content of the video and N_v is the set of not-verified tags - that couldn't be identified as relevant to the content of the video.

4. PROPOSED APPROACH

In this section, we provide a detailed description of the proposed approach for identifying the relevance of user generated tags with respect to the content of the video. This also allows us to detect irrelevant tags and misleading metadata for online videos. An overview of the approach in described in figure 1. The input to the algorithm is the video content, tags and comments on the video. We process user generated content by tag processing and analyze videos to extract features and then use them to verify certain tags. Then, we use comments for tag verification depending on their overall cumulative sentiment.

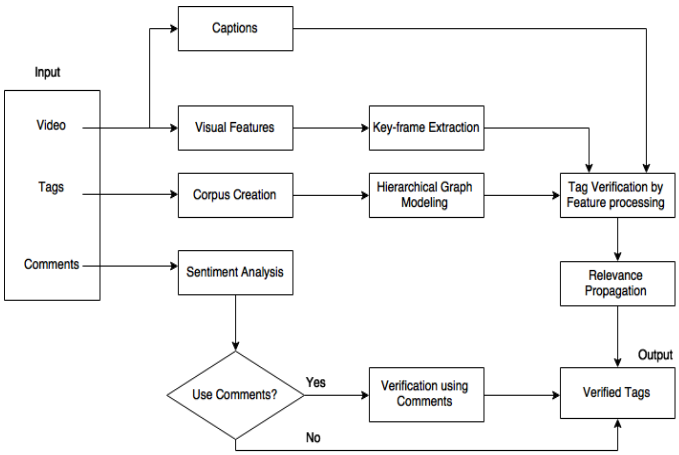

Figure 1: Proposed Approach

4.1 Tag Processing

The algorithm begins by analyzing the tags provided by the users. The type and variations of these tags makes it difficult to use them directly for identifying relevance with video content. This is because of the following reasons (1) Redundancy: Users tend to associate all variants of the tags with the video (2) Abbreviations: Users tend to use abstractions for popular terms (3) Polysemy: Tags can be ambiguous with multiple inferences (4) Compound Terms and Mixed Entities: Users split multi-word entities, which if taken separately might have completely different meanings, and also combine multiple meaningful entities together, which might make more sense if split appropriately. (5) Tag Variations: Tags can very specific like conceptual models (*basketball*) and very generic and abstract(*movie*).

Therefore, there is a need to pre-process the tags to address these issues before the verification process begins. We do this by relating tags with respect to the content and context of the video, because context modeling helps to understand the meaning of tags when it is used with other tags. For example, the tag *matrix* might correspond to algebra

or it might refer to the movie. Relating it to other tags provided by the users can help disambiguate the context for the tag. For example, if the tag *actor* is present with the tag *matrix*, it makes it possible to disambiguate the latter. The algorithm takes in the set of tags T_v and title H_v for every video and outputs a hierarchical graph G_v for the video v in the set V. The approach of tag processing is represented in algorithm 1.

4.1.1 Corpus Creation

The method begins with gathering the semantics of all the tags a video. We create a corpus for each tag in user provided metadata by using the unstructured web content of the top ten search results when this tag and video title are queried on a search engine. Corpus creation using web mining has been well studied in information retrieval. Web mining allows us to capture all possible meanings of the tags [32]. We extract this content using Bing API [38]. In algorithm 1, a corpus C_v represents the corpus for video v.

4.1.2 Hierarchical Graph Modeling

After capturing tag semantics, we use unstructured data from the Web to tackle the problem of context resolution using tag associations We bridge the gap between user assigned keywords for establishing tag correlations by creating a concept hierarchy of tags for every video. This makes it possible to exploit pairwise meaningful relationships between the tags and to handle the level of generality in the tags uploaded by the users. For instance, *bat* might not be directly related to *movie*, but this link can be completed with the assistance of other tags - *bat* is related to *batman*, and *batman* is related to *movie*. Thus, hierarchical modeling enables us to handle all varieties of generality that the users might use in their metadata and tags. In this way, the relationships assess the meaning of tags with respect to other tags thus handling ambiguities and polysemy, and graph modeling addresses all the variations by creating hierarchical parent-child edges between these tags.

Taxonomy construction techniques are well studied and the major approaches are based on clustering, lexico-syntactic pattern matching, automatic acquisition from machine-readable dictionaries and heuristic rules and deep syntactic analysis [33]. Here, we provide a simplistic approach to determine the hierarchical semantic relationships between tags by extending the concept of Google distance [12] which measures co-occurrence in text search results and computes similarity between words using Google page counts.

We begin by initializing a graph G_v with the tags as nodes and no edges. Following this, each pair of tags is analyzed for a potential edge in the graph depending on corpus co-occurrence. We empirically calculate pairwise conditional probabilities. This approach is similar to taxonomy construction in [5]. $p(t_1|t_2)$ is the number of search results in which tag t_1 occurs in the corpus of tag t_2. If the difference in conditional probabilities $p(t_1|t_2)$ and $p(t_2|t_1)$ exceeds α, which is a user defined threshold, an edge is created from t_2 to t_1, i.e., t_2 is a generic tag and t_1 is a specific tag. Therefore, t_1 becomes the child of t_2. For instance, tag t_1 is *movie* and tag t_2 is *avengers*, then $p(t_1|t_2)$ would be very large compared to $p(t_2|t_1)$, because the search results for *movie* may or may not have the tag *avengers*, but the search results for *avengers* will definitely have the tag *movie* in them. So if $p(t_1|t_2) - p(t_2|t_1) > \alpha$, then an edge is created from t_1 to t_2. We discuss the impact of choice of α later in the experiments section.

Algorithm 1 Pseudo code of Hierarchical Graph Modeling

Input: $V, \{T_v, H_v\} \forall v \in V$
Output: $G_v \forall v \in V$
for each *video* $v \in V$ **do**
  Corpus Creation
  $C_v = \{\}$
  for each *tag* $t \in T_v \cup H_v$ **do**
    $C_t = Query(t)$
    $C_v = C_v \cup C_t$
  end for
  Hierarchical Graph Modeling
  $G_v = (T_v, \phi)$
  for each *tag* $t_1 \in T_v$ **do**
    for each *tag* $t_1 \in T_v$ *and* $t_1 \neq t_2$ **do**
      $p(t_1|t_2) = |t_1 \in C_{t_2}|/|C_{t_2}|$
      $p(t_2|t_1) = |t_2 \in C_{t_1}|/|C_{t_1}|$
      if $(p(t_1|t_2) - p(t_2|t_1) > \alpha)$ **then**
        $E(G_v) = E(G_v) \cup edge(t_2 -> t_1)$
      end if
    end for
  end for
  $G_v = Transitive\_Reduction(G_v)$
end for

4.1.3 Entity Extraction

The tags still might not be representative enough to be able to deduce any meaningful relationships between them. For instance, tags such as *basketball* and *ring* are three separate tags in uploader provided metadata but they can be combined into an entity in context of the video. The tag *ring* can have multiple meanings when considered independent of the tag *basketball*. Therefore, we extract meaningful entities from the tags by combining them which helps us to handle redundancy, abbreviations and compound terms. We exploit the pairwise semantic relationships between tags by leveraging conditional probabilities in the graph G_v: for every pair of tags t_1 and t_2 connected via an edge in graph G_v, if both $p(t_1|t_2)$ and $p(t_2|t_1)$ are both greater than 0.8, the tags corresponding to the same entity are combined into the same node in the graph. Thus, very high correlation between tags is used to combine redundant tags, connect abbreviations with corresponding complete terms and fuse multiple terms to extract meaningful entities which can then be verified through further processing.

4.2 Feature Processing

In this section, we describe the features that are extracted from the videos for tag verification. We primarily rely on the textual features provided as captions and video features extracted as image keyframes.

4.2.1 Textual Features

For every video, we use the captions associated with the video as the speech content belonging to the video. Entities are extracted from these captions are used to verify the relevance of some of the entities in the graph. It might be possible that a user who gives misleading tags in the metadata uses those incorrect tags in the captions as well. To avoid this, we only consider the captions which are auto-generated

by YouTube which are based on Google's automatic speech recognition (ASR) technology ([28]).

4.2.2 Video Features

Video feature processing is a crucial part of the algorithm as here we verify that the actual content of the video matched the user generated tags for that video. We extract keyframes from the video using ([26], [29]) by detecting abrupt and gradual transitions, based on the visual similarity of neighboring frames of the video. We exploit the descriptive efficiency of both local (SURF) and global (HSV histograms) descriptors to assess frame similarity. Specifically, abrupt transitions are initially detected between successive video frames where there is a sharp change in the visual content, which is expressed by a very low similarity score. Then, we analyze the calculated scores to identify frame-sequences where a progressive change of the visual content takes place to detect gradual transitions. Finally, we identify and remove outliers due to object/camera movement and flash-lights.

We verify the entities extracted from the tags through tag processing in the extracted keyframes by using the overfeat classifier [31]. The OverFeat convolutional net is trained on the ImageNet 1K dataset. Each image is downsampled so that the smallest sample is 256 pixel. Random crops of size 221x221 are then presented to the network. The first two layers each involve a convolutional stage followed by a max-pooling stage. 3rd and 4th involve only convolution. Layer 5 is the top convolution plus max pooling layer. Subsequent layers are fully connected. The network has eight layers with the final layer having dimensions 1000x1.

The tags we validate are restricted to those present in ImageNet, and can be easily extended to more tags using other image classifiers. For instance, there has been significant work in extending the trained Convolutional Network Overfeat to move further away from the original task and the data that the OverFeat was trained to solve [30].

For celebrity verification, we first detect and extract the faces using Haar Cascades from every keyframe extracted, align them [34] and then tag them. We ignore the tags whenever the confidence for the predicted tag was below 80%. It is very important for the face recognition module to have very few false positives.

4.3 Relevance Propagation

In this section, we prose a novel approach to assign relevance to tags which couldn't directly be identified by feature processing. Specifically, generic and abstract tags like *movie*, *game*, etc. can't be verified by image similarity techniques. We use the hierarchical graph created using algorithm 1 for this purpose. The method we follow is summarized in algortihm 2. The algorithm takes as input the hierarchical graph G_v and the set of tags verified through feature processing V_v, and it returns the set V_v, that is the set of tags identified as related to the content of the video. We initialize an empty set V_v, and the repeated the following until V_v becomes stable, that is, no other tags are added to it: we analyze the tags in the parent nodes for all the verified and related tags in V_v respectively, and if these tags are not already in V_v, they are added to V_v.

Figure 2 shows an example of a hierarchical graph of the tags generated using our algorithm for a video taken from *YouTube*. Note that all the semantically related entities-

Algorithm 2 Pseudo code of Relevance Propagation

Input: $\{T_v, G_v, V_v\}\ \forall\ v \in V$
Output: $V_v\ \forall\ v \in V$

```
for each video v ∈ V do
  while V_v is not stable do
    for each tag t_1 ∈ V_v do
      P_{t_1} = ∪{t_2 | t_2 ∈ T_v, edge(t_1, t_2) ∈ Edges(G_v)}
      for each tagt_2 ∈ P_{t_1} do
        if (t_2 ∉ {V_v}) then
          V_v = V_v ∪ t_2
        end if
      end for
    end for
  end while
end for
```

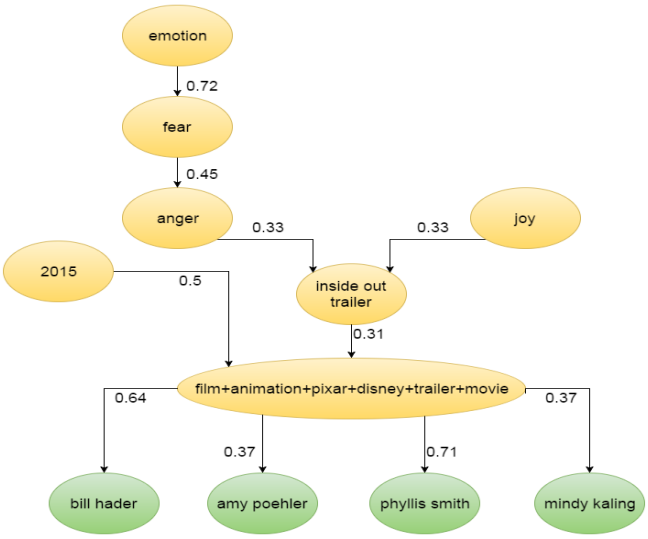

Figure 2: Example: Tag Processing

film, *animation*, *pixar*, *disney*, *trailer* and *movie* have been clubbed together. Redundancies in the tags *new trailers*, *trailers HD*, and *trailer* have been reduced and meaningfully related to the entity *trailer*. This demonstrates the impact of establishing tag relationships and entity extraction. The tag 2015 has multiple perceptions, but it's semantics in context is extracted through graph modeling. Without such meaningful relationships, it is difficult to check the relevance of such tags.

Also, celebrities in the movie cast are the children of the clubbed entity in the graph. The verification of these celebrities through face detection (marked in green) assist in verify other tags up in the hierarchy through relevance propagation, which are marked in yellow. Tags such as *fear*, *anger*, and *joy*, which are difficult to detect through image feature processing can be verified through relevance propagation. This also leads to verification of the tag *emotion*.

4.4 Comment Analysis

Comments can be a useful source of information about the relation between the title and content of the video, because it is mostly the title and the thumbnail that governs the selection of the video after search, and users tend to express their sentiments in the comments if they are disappointed

by misleading metadata, i.e., when the content of the video is different from what they had expected. We introduce the concept of *Cumulative Sentiment* to use comments for tag verification. For every video v *in* V, let the number of comments be c_v. We define the cumulative sentiment as

$$CS_v = \sum_{i=1}^{c_v} Sentiment(c_i) \quad (2)$$

Here, we use polarity of the comments ([23]) to calculate sentiment of a comment which uses natural language processing, text analysis and computational linguistics to identify and extract subjective information in text. The value $Sentiment(c_i)$ lies between -1 to 1. We take top 500 comments for every video, so the value of cumulative sentiment can range anywhere between -500 to 500.

We assume that an overall increasing nature of the cumulative sentiment over time indicates that the title pertains to the content of the video. If this holds for a video, i.e., the overall cumulative sentiment is positive, we form the title corpus by first querying the title over the web and then extracting text from the top ten search results. We make sure that links to the video are not used in the corpus. We use this corpus to verify entities. This is because positive sentiment reflects that the visible metadata, that is, the title of the video is agreeing with the content of the video and since the visible metadata is the title we use information related to it to verify the tags. This allows us to incorporate external information from the web for tag validation.

5. EXPERIMENTS

In this section, we present our experimental results on assessing the relevance of user generated tags.

5.1 Dataset

The experiment was performed on videos and corresponding user assigned metadata from the popular video hosting website $YoutTube$. We collected about 300 videos and associated title, metadata, description and tags using $YouTube\ API$[37]. We also extract the captions and comments for the videos when available. The length of the videos range from 2 to 10 minutes. Also, we used 100 videos from three different video categories - $Music, Entertainment$ and $Movie$ to evaluate our algorithm.

5.2 Ground Truth

Videos collected for the experiments were distributed across many users for manual annotation. These users examined the videos and inspected user generated tags with respect to content of the video. First, these users classified the tags into two categories - *specific* and *generic*, where former refers to the objects and faces which Then, they classified the tags into two categories, i.e., if they are relevant to the video content or not. To make sure that the ground truth is not biased by a users' perceptions, we take at least five users to annotate every tag for all videos, and the ground truth is taken as the maximum response of all the users.

5.3 Exploratory Data Analysis

In this section, we perform an initial analysis on the dataset to understand the nature of the videos and associated tags. First, we provide an analysis on the number of tags for the collected videos. Figure 3 shows this number ranges from 0 to about 60. High number of tags is not a surprise because the metadata associated with the videos directly impacts the rank of the video in the search results. Further, the number of tags vary with video category: it can be seen that users tend to use many tags for *Entertainment* and *Movie* as compared to the number of tags for *Music*. 90% of *Music* videos have less than 25 tags whereas 95% of *Entertainment* videos have less than 55 tags. Most of the *Movie* videos have 20 to 35 tags.

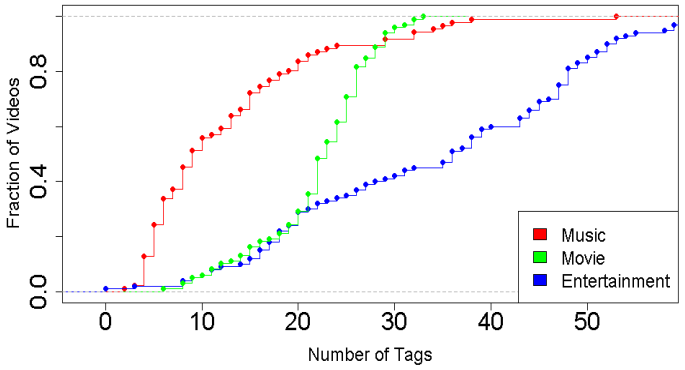

Figure 3: Number of Tags by Video Category

Next, we analyze the variety of tags assigned by the users. Here, the generic tags refers to the tags which can't be identified by automatic image annotation or techniques. As it can be seen in figure 4, the number of generic tags are many more than specific tags for each category. Note that the number of generic and abstract tags are as high as 95% for all the three categories. Therefore, if we just use image feature processing techniques for checking the relevance of tags with respect to video content, we won't be able to verify majority of the tags even if they are not misleading, thus leading to very low recall values. We discuss this further in the results section.

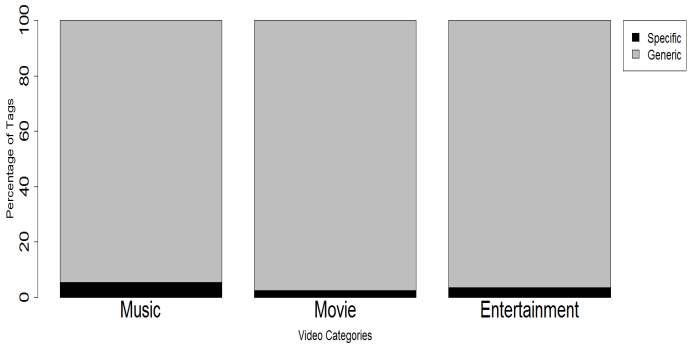

Figure 4: Generic vs. Specific Tags

5.4 Choosing Parameter α

In this section, we describe the choice of α for our experiments. This threshold governs the level of connectivity and generality which is acceptable in the hierarchical semantic graph G_v. Basically, α governs if the difference in the conditional probabilities $p(t_1|t_2)$ and $p(t_2|t_1)$ for the pairs of tags

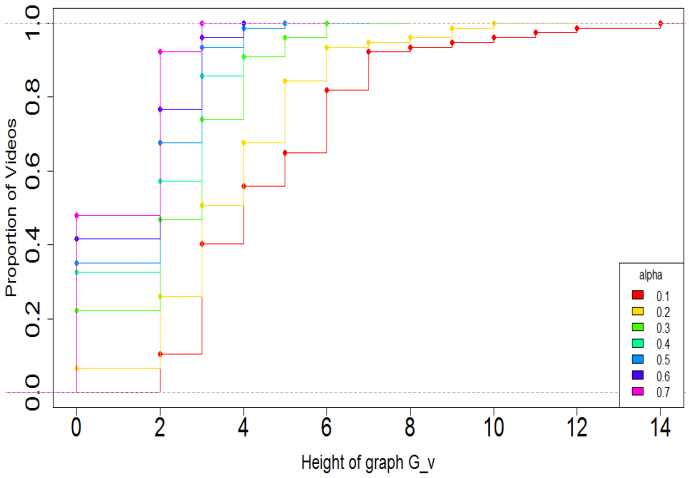

Figure 5: Cumulative Distribution of height of G_v for different values of α

$t1$ and $t2$ is high enough to create edges between them in the graph. If the value of α is too high, it requires the values of conditional probabilities to have a high difference in order to create an edge between them. This constraint would result in the graph with have very few edges and smaller height after transitive reduction. On the other side, if α is low, the constraint on difference in conditional probabilities is slackened, thus allowing more edges in the graph resulting in longer heights for the graphs. But if the value of α becomes too low, it might result in connecting many unrelated tags in the graph and causing the height to increase a lot.

This conjecture can be verified using fig 5. This plot demonstrates the cumulative distribution of the heights of the hierarchical graphs for tags of videos in consideration for different values of α. Note that when the values of α is 0.1 or 0.2, we get many graphs with height of 6 or more, because a lower threshold allows too many edges in the graph causing the height to increase. And as the value of α increases beyond 0.2, we get fewer edges and thus smaller heights. Note if the values of α goes beyond 0.3, we start getting a lot of graphs with very less height. For instance, when the values of α is 0.3, about 20% of the graphs have height 2 or less, and this number increases up to nearly 35% and 40% when α is increased to 0.5 and 0.6. We notice that values of α beyond 0.3 and 0.4 misses out on capturing essential relations between tags because the high constraint results in too many isolated tags. For instance, when α is between 0.5 and 0.7, 70% to 90% of the videos have height less than 3. This creates a problem for our relevance proportion algorithm which relies on propagating the relevance from specific to generic tags because many generic tags get isolated from specific tags. Therefore, we choose the value of α to be 0.3.

6. RESULTS

In this section, we demonstrate the experimental results of our algorithm on the dataset. First, we describe the process of identifying which tags are relevant to the content of the video and then evaluate our results against the ground truth.

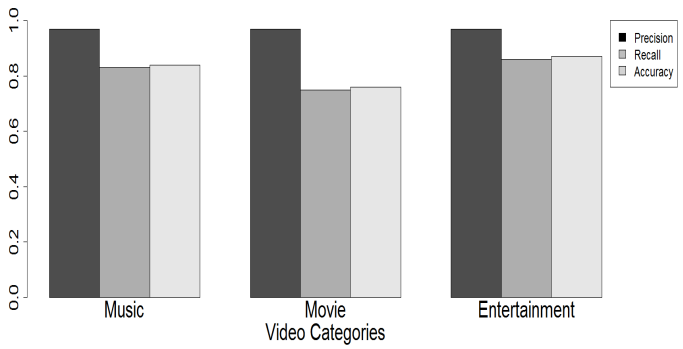

Figure 6: Precision, Recall and Accuracy by Video Category

Category		Relevant	Irrelevant
Music	Predicted Relevant	611	15
	Predicted Irrelevant	121	107
Movie	Predicted Relevant	1361	44
	Predicted Irrelevant	408	68
Entertainment	Predicted Relevant	1174	36
	Predicted Irrelevant	192	464

Table 1: Confusion Matrix

6.1 Experimental Evaluation

We evaluate the performance of our algorithm by comparing it against the ground truth. The tags which are classified as to be in the verified set V_v are predicted to be relevant to the content of the video by our algorithm, where as the tags in the not verified set N_v are taken to be irrelevant to the content of the video. The confusion matrices for our results are presented in table 1. We analyze the efficiency of the algorithm in terms of *Precision*, *Recall* and *Accuracy* values:

$$Precision = \frac{\sum_{v \in V} |Correctly\ classified\ Relevant\ tags\ for\ v|}{\sum_{v \in V} |Tags\ classified\ as\ Relevant\ for\ v|} \quad (3)$$

$$Recall = \frac{\sum_{v \in V} |Correctly\ classified\ Relevant\ tags\ for\ v|}{\sum_{v \in V} |Relevant\ tags\ for\ v|} \quad (4)$$

$$Accuracy = \frac{\sum_{v \in V} |Correctly\ classified\ tags\ for\ v|}{\sum_{v \in V} |Total\ number\ of\ tags\ for\ v|} \quad (5)$$

As figure 6 shows, we are able to detect majority of the relevant tags as we achieve high recall and precision values for all three categories. The recall for *Music*, *Movie* and *Entertainment* are 0.83, 0.75 and 0.86 and precision for all these three categories is about 0.97. The accuracy values for *Music*, *Movie* and *Entertainment* are 0.84, 0.76 and 0.87. High recall numbers demonstrate that we verify majority of the relevant metadata for the video while achieving very high precision values, thus with very few false positives.

We achieve comparatively lower recall values for *Movie* category as compared to the other two categories because

there are several tags which are related to the video but may not be always present in the video content. This problem surfaces strongly in case of *Movie* videos because there are many such tags in this category, such as the director, other off-screen crew members or may be some actor who has a brief role in the movie. These will have a less possibility to have any appearance in it. And as the relationships built are hierarchical, relevance is never propagated to proper nouns if they form the leaf nodes in the graph. Therefore, these tags will not be identified by our algorithm.

We also compare the performance of our method against using image feature processing only. In order to do this, we use image annotation and face recognition techniques to identify relevant tags with respect to content of the video. We get better recall of 0.813 as compared to 0.10 obtained using just image feature processing techniques . Though image annotation and face recognition yield he perfect precision value, these methods fails to recognize majority of user provided tags. It is because the user-assigned tags consist of a large number of generic tags as already shown above, and it is difficult to identify such tags using image processing standalone. Through the proposed method, we are able to achieve higher recall values, while keeping precision values close to 1. The maximum improvement is for *Entertaiment* category where recall increases from 0.06 using image processing to 0.86 using our method, as we were able to verify many generic and abstract tags in this category.

6.2 Incremental Performance Analysis

In this section, we analyze the incremental impact of each piece of the algorithm - Using just image processing, adding hierarchical graph modeling and relevance propagation and then comment analysis for tag verification. Our graph modeling and relevance propagation algorithm improves accuracy from 0.08 using just image processing to 0.27. This is because leveraging semantic tag relationships enables us to detect many generic and abstract tags which are relevant to the video content and which are very difficult to be identified using conventional image processing techniques.

Further, using comments along with graph modeling and relevance propagation increases accuracy to about 0.80 aggregated over all three categories.

6.3 Fraudulent Metadata and Irrelevant Tags

We performed an additional analysis on the number of irrelevant metadata added by the users for videos of all three categories. We have manually identified misleading tags for all the videos and compared the accuracy of our system in terms of identifying such tags. The results are summarized in figure 7. Note that, the proposed approach is able to detect such tags with an accuracy of approximately 90% for *Music* and *Entertainment* where the proportion of such tags is higher comparative to *Movie*, where approximately 9% of the tags are irrelevant and the accuracy of our system drops to around 60%.

7. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel problem of identifying the relevance of use generated metadata for online videos. This can also be used for detecting irrelevant user generated tags and misleading metadata. The variation in user generated tags makes the problem hard. We also presented a novel approach to solve this problem by modeling hierarchical semantic relationships between tags and using audiovisual features of the video to identify relevant specific tags. We then introduced a relevance propagation technique to identify the relevance of generic and abstract tags. We also presented an analysis of how sentiment analysis on comments can be leveraged for tag relevance identification for *YouTube* videos. Further, we demonstrated the results of our algorithm on 300 *YouTube* videos.

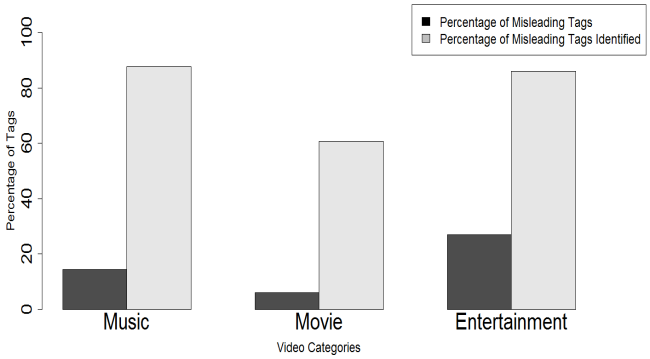

Figure 7: Misleading Tags by Video Category

Our algorithm relies on the corpus created by web mining. One direction of future work is to improve the quality of search results to ensure stability of our results. The next stage of our research would be to increase the size of our dataset and understand the impact of misleading metadata in more videos beyond the three categories presented in this paper. At the same time, we can perform more experiments to analyze users who upload videos with misleading metadata. Trend analysis for the users who continuously upload irrelevent content or the topics for which incorrect tags are posted can be used to impose appropriate restrictions for online media platforms. Total relevance of the tags can be used to generate better search results and to measure the cohesiveness of content are possible directions of future work.

8. REFERENCES

[1] Miller, George A., et al. "Introduction to wordnet: An on-line lexical database*." International journal of lexicography 3.4 (1990): 235-244.

[2] Benevenuto, Fabricio, et al. "Practical detection of spammers and content promoters in online video sharing systems." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42.3 (2012): 688-701.

[3] Bulakh, Vlad, Christopher W. Dunn, and Minaxi Gupta. "Identifying fraudulently promoted online videos." Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014.

[4] Li, Haojie, et al. "DUT-WEBV: a benchmark dataset for performance evaluation of tag localization for web video." Advances in Multimedia Modeling. Springer Berlin Heidelberg, 2013. 305-315.

[5] Liu, Xueqing, et al. "Automatic taxonomy construction from keywords." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[6] Fogg, B. J., et al. "What makes Web sites credible?: a

report on a large quantitative study." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2001.

[7] Toderici, George, et al. "Finding meaning on youtube: Tag recommendation and category discovery." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[8] Aradhye, Hrishikesh, George Toderici, and Jay Yagnik. "Video2text: Learning to annotate video content." Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on. IEEE, 2009.

[9] Nguyen, Nam P., et al. "Containment of misinformation spread in online social networks." Proceedings of the 4th Annual ACM Web Science Conference. ACM, 2012.

[10] Adler, B. Thomas, and Luca De Alfaro. "A content-driven reputation system for the Wikipedia." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[11] Tambuscio, Marcella, et al. "Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks." Proceedings of the 24th International Conference on World Wide Web. ACM, 2015.

[12] Cilibrasi, Rudi L., and Paul MB Vitanyi. "The google similarity distance." IEEE Transactions on knowledge and data engineering 19.3 (2007): 370-383.

[13] Stefan Siersdorfer, Jose San Pedro, and Mark Sanderson. 2009. Automatic video tagging using content redundancy. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)

[14] Ting Yao, Tao Mei, Chong-Wah Ngo, and Shipeng Li. 2013. Annotation for free: video tagging by mining user search behavior. In Proceedings of the 21st ACM international conference on Multimedia (MM '13)

[15] Wathen, C. Nadine, and Jacquelyn Burkell. "Believe it or not: Factors influencing credibility on the Web." Journal of the American society for information science and technology 53.2 (2002): 134-144.

[16] Li, Li-Jia, Richard Socher, and Li Fei-Fei. "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[17] Li, Xirong, Cees GM Snoek, and Marcel Worring. "Learning tag relevance by neighbor voting for social image retrieval." Proceedings of the 1st ACM international conference on Multimedia information retrieval. ACM, 2008.

[18] Li, Xirong, Cees GM Snoek, and Marcel Worring. "Learning tag relevance by neighbor voting for social image retrieval." Proceedings of the 1st ACM international conference on Multimedia information retrieval. ACM, 2008.

[19] Li, Xirong, Cees GM Snoek, and Marcel Worring. "Learning tag relevance by neighbor voting for social image retrieval." Proceedings of the 1st ACM international conference on Multimedia information retrieval. ACM, 2008.

[20] Jeon, Jiwoon, Victor Lavrenko, and Raghavan Manmatha. "Automatic image annotation and retrieval using cross-media relevance models." Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003.

[21] Carneiro, Gustavo, et al. "Supervised learning of semantic classes for image annotation and retrieval." Pattern Analysis and Machine Intelligence, IEEE Transactions on 29.3 (2007): 394-410.

[22] Jeon, Jiwoon, and R. Manmatha. "Using maximum entropy for automatic image annotation." Image and Video Retrieval. Springer Berlin Heidelberg, 2004. 24-32.

[23] https://pypi.python.org/pypi/textblob-de/

[24] Sargin, Mehmet Emre, et al. "Audiovisual celebrity recognition in unconstrained web videos." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.

[25] Zhao, John, et al. "Large scale learning and recognition of faces in web videos." Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on. IEEE, 2008.

[26] E. Apostolidis, V. Mezaris, "Fast Shot Segmentation Combining Global and Local Visual Descriptors", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 2014.

[27] Blumenstock, Joshua E. "Size matters: word count as a measure of quality on wikipedia." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

[28] https://googleblog.blogspot.in/2009/11/automatic-captions-in-youtube.html

[29] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, I. Trancoso, "Temporal video segmentation to scenes using high-level audiovisual features", IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 8, pp. 1163-1177, August 2011.

[30] Razavian, Ali S., et al. "CNN features off-the-shelf: an astounding baseline for recognition." Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. IEEE, 2014.

[31] Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229 (2013).

[32] Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the Web." (1999).

[33] Tsui, Eric, et al. "A conceptâĂŞrelationship acquisition and inference approach for hierarchical taxonomy construction from tags." Information processing & management 46.1 (2010): 44-57.

[34] Huang, Gary B., Vidit Jain, and Erik Learned-Miller. "Unsupervised joint alignment of complex images." Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007.

[35] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. "Yago: a core of semantic knowledge." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[36] https://www.youtube.com/yt/press/statistics.html

[37] https://developers.google.com/youtube/v3/?hl=en

[38] https://datamarket.azure.com/home