FunTube: Annotating Funniness in YouTube Comments

Laura Zweig, Can Liu, Misato Hiraga, Amanda Reed,
Michael Czerniakowski, Markus Dickinson, Sandra Kübler

Indiana University
{lhzweig,liucan,mhiraga,amanreed,emczerni,md7,skuebler}@indiana.edu

1 Introduction and Motivation

Sentiment analysis has become a popular and challenging area of research in computational linguistics (e.g., [3, 6]) and even digital humanities (e.g., [10]), encompassing a range of research activities. Sentiment is often more complicated than a positive/neutral/negative distinction, dealing with a wider range of emotions (cf. [2]), and it can be applied to a range of types of text, e.g., on YouTube comments [9]. Sentiment is but one aspect of meaning, however, and in some situations it can be difficult to speak of sentiment without referencing other semantic properties. We focus on developing an annotation scheme for YouTube comments, tying together comment relevance, sentiment, and, in our case, humor.

Our overall project goal is to develop techniques to automatically determine which of two videos is deemed funnier by the collective users of YouTube. There is work on automatically categorizing YouTube videos on the basis of their comments [5] and on automatically analyzing humor [4]. Our setting is novel in that for YouTube comments each comment does not necessarily itself *contain* anything humorous, but rather points to the humor within another source, namely its associated video (bearing some commonality with text commentary analysis, e.g., [11]).

For our annotation of user comments on YouTube humor videos, a standard binary (+/-) funny annotation would ignore many complexities, stemming from different user motivations to leave comments, none of which include explicitly answering our question. We often find comments such as *Thumbs up for Reginald D. Hunter!* (`https://www.youtube.com/watch?v=tAwZL3n9kGE`), which is clearly positive, but it is unclear whether it is about funniness.

We have developed a multi-level annotation scheme (section 3) for a range of video types (section 2) and have annotated user comments for a pilot set of videos. We have also investigated the impact of annotator differences on automatic classification (section 5). A second contribution of this work, then, is to investigate the connection between annotator variation and machine learning outcomes, an important step in the annotation cycle [8] and in comparing annotation schemes.

2 Data Collection

Our first attempt to extract a set of funny videos via the categories assigned by the video uploader failed, as many videos labeled as `comedy` were mislabeled (e.g., a `comedy` video of a train passing through a station). Thus, we started a collection with a diversity of different categories, covering different types of humor, and began gathering this data semi-automatically. To ensure broad coverage of different varieties, we formed a seed set of 20 videos by asking for videos from family, friends, and peers. We asked for videos that: a) they found hilarious; b) someone said was hilarious but was not to them (or vice versa); and c) "love it or hate it" types of videos.[1] The seed set covers videos belonging to the categories: `parody`, `stand-up`, `homemade`, `sketch`, and `prank`. We used the Google YouTube API (`https://developers.google.com/youtube/`) to obtain 20 related videos for each seed (˜100 total videos), filtering by the YouTube `comedy` tag. The API sorts comments by the time they were posted, and we collected the 100 most recent comments for each video (˜10,000 total comments). In case of conversations (indicated by the Google+ activity ID), only the first comment was retrieved. Non-English comments were filtered via simple language identification heuristics.

3 Annotation Scheme

Intuitively, annotation is simple: For every comment, an annotation should mark whether the user thinks a video is funny or not—perhaps allowing for neutral or undetermined cases. But consider, e.g., a sketch comedy video concerning Scottish accents (`http://www.youtube.com/watch?v=BncDeMO_en0`). While there are obvious cases for annotation like "LOL" and "these shit is funny", there are cases such as in (1) that are less clear.

(1) a. I think British Irish and Scottish accents are the most beautiful accents.
b. Its "Burnistoun" on The BCC One Scotland channel! It's fantastic.
c. ALEVENN

In (1a) the user is expressing a positive comment, which—while related to the video—does not express any attitude or information concerning the sketch itself. In (1b), on the other hand, the comment is about the video contents, informing other users that the clip comes from a show *Burnistoun* and that the user finds this show to be fantastic. Two problems arise in classifying this comment: 1) it is about the general show, not this particular clip, and 2) it expresses positive sentiment, but not directly about humor. Example (1c) shows the user quoting the clip, and, as such, there again may be an inference that they have positive feelings about the video, possibly about its humor. The degree of inference that an annotator should

[1] There will be some bias in this method of collection, as humor is subjective and culturally specific; by starting from a diverse set of authors, we hope this is mitigated to some extent.

Level	Labels									
Relevance	**R(elevant)**								I(rr.)	U(nsure)
Sentiment	**P(ositive)**			**N(egative)**		Q(uote)	A(dd-on)	U(nclear)		
Humor	**F(unny)**	N(ot Funny)	U(nclear)	**F**	N U					

Table 1: Overview of the annotation scheme

draw during the annotation process must be spelled out in the annotation scheme in order to obtain consistency in the annotations. For example, do we annotate a comment as being about funniness only when this is explicitly stated? Or do we use this category when it is reasonably clear that the comment implies that a clip is funny? Where do we draw the boundaries?

Funniness is thus not the only relevant dimension to examine, even when our ultimate goal is to categorize comments based on funniness. We account for this by employing a tripartite annotation scheme, with each level dependent upon the previous level; this is summarized with individual components of a tag in table 1. The details are discussed below, though one can see here that certain tags are only applicable if a previous layer of annotation indicates it, e.g., the `F` funniness tag only applies if there is sentiment (`P`ositive or `N`egative) present. For examples of the categories and of difficult cases, see section 4.

This scheme was developed in an iterative process, based on discussion and on disagreements when annotating comments from a small set of videos. Each annotator was instructed to watch the video, annotate based on the current guidelines, and discuss difficult cases at a weekly meeting. We have piloted annotation on approx. 20 videos, with six videos used for a machine learning pilot (section 5).

3.1 Relevance

First, we look at relevance, asking annotators to consider whether the comment is relevant to the *contents* of the video, as opposed to side topics on other aspects of the video such as the cinematography, lighting, music, setting, or general topic of the video (e.g., homeschooling). Example (2), for instance, is not relevant in our sense. Note that determining the contents is non-trivial, as whether a user is making a specific or general comment about a video is problematic (see section 4 for some difficult cases). By our definition, the contents of the video may also include information about the title, the actors, particular jokes employed, dialogue used, etc. To know the relevance of a comment, then, requires much knowledge about what the video is trying to convey, and thus annotators must watch the videos; as discussed for (1c), for example, the way to know that references to the number 11 are relevant is to watch and note the dialogue.

(2) This video was very well shot

Turning to the labels themselves, annotators choose to tag the comment as `R` (relevant), `I` (irrelevant), or `U` (unsure). As mentioned, since relevance is based on

the content of the video, comments about the topic of the video are not considered relevant. Thus, the comment in (3) is considered irrelevant for a video about homeschooling even though it does refer to homeschooling. Only if the comment receives an `R` tag does an annotator move on to the sentiment level.

(3) Okay, but homeschooling is not that bad!

Relevance—and particularly relevance to the video's humor—is a complicated concept [1]. For one thing, comments about an actor's performance are generally deemed relevant to the video, as people's impressions of actors are often tied to their subjective impression of a video's content and humor. In a similar vein regarding sentiment, general *reactions* to the contents of the video are also considered relevant, even if they do not directly discuss the contents of the video. Several examples are shown in (4). Note that in all cases, the video's contents is the issue under discussion in these reactions.

(4) a. I love Cracked!
b. The face of an angel! lol
c. This is brilliant!

One other notion of relevancy is concerned with the idea of interpretability and whether another user will be able to interpret the comment as relevant. For example, all non-English comments are marked as irrelevant, regardless of the content conveyed in the language they are written in. Likewise, if the user makes a comment that either the annotator does not understand or thinks no one else will understand, the comment is deemed irrelevant. For example, a video of someone's upside-down forehead (bearing a resemblance to a face) generates the comment in (5). If we tracked it down correctly, this comment is making a joke based on a reference to a 1993 movie (*Wayne's World 2*), which itself referenced another contemporary movie (*Leprechaun*). Even though it references material from the video, the annotator assumed that most users would not get the joke.

(5) I'm the Leprechaun.

3.2 Sentiment

Sentiment measures whether the comment expresses a positive or negative opinion, regardless of the opinion target. Based on the assumption that quotes from the video make up a special case with unclear sentiment status (cf. (1c)), annotators choose from: `P` (positive), `N` (negative), `Q` (quote), `A` (add-on), and `U` (unclear). `Q` is used for direct quotes from the video (without any additions), and `U` is used for cases where there is sentiment but it is unclear whether it is positive or negative.

For example, in (6), the user may be expressing a genuine sentiment or may be sarcastic, and the annotation is `U`. In general, in cases where the comment does not fit any of the other labels (`P`, `N`, `Q`, `A`), the annotator may label the comment as `U`.

(6) This is some genius writing

A is used in cases where the comment responds to something said or done in the video, usually by attempting to add a joke by referencing something in the video. An example comment tagged as A is shown in (7), which refers to a question in the homeschooling video about the square root of 144. Again, note that add-ons, like quotes, require the annotator to watch and understand the video, and, as mentioned in section 3.1, add-ons are only included if the add-on is clearly understandable.

(7) It's 12!

The positive and negative cases are the ones we are most interested in, for example, as in the clear case of positive sentiment in (8). Only labels of P and N are available for the final layer of annotation, that of humor (section 3.3). If an annotator cannot reasonably ascertain the user's sentiment towards the video, then it is unlikely that they will be able to determine the user's feelings about the humor of the video. In that light, even though Q and A likely suggest that the video is humorous, we still do not make that assumption.

(8) This was incredible! I'm sooo glad I found this.

In terms of both relevance and sentiment, we use a quasi-Gricean idea: If an annotator can make a reasonable inference about relevance or sentiment, they should mark the video as such. In (9), for instance, the comment refers to a part of the video clip where the student sarcastically comments that he will go to "home college." Thus, it seems reasonable to make an inference about the sentiment.

(9) I graduated home college!

3.3 Humor

With the comment expressing some sentiment, annotators then mark whether the comment mentions the funniness of the video (F) or not (N)—or if it is unclear (U). In this case, we are stricter about the definition: if it is not clearly about funniness, it should not be marked as such. For example, the different comments in (10) are all relevant (R) and positive (P), but do not specifically refer to the humor and are marked as N. In general, unless the comment explicitly uses a word like "funny" or "humor", it will likely be labeled as N or U.

(10) a. This is the most glorious video on the internet.
b. This is brilliant!
c. This is some genius writing!

Note that by the time we get to the third and final layer of annotation, many preliminary questions of uncertainty have been handled, allowing annotators to focus only on the question of whether the user is commenting on the video's humor.

4 Examples

We present examples for cases that can easily be handled by an annotation scheme and others that are more difficult. The latter indicate where guidelines need refinement and where automatic systems may encounter difficulties.

Clear Cases Consider the comment in (11a), annotated as `RPF`: The comment directly mentions the video (`R`), has positive sentiment (`P`), and directly comments on the humor of the video, as indicated by the word "hilarious" (`F`). The comment in (11b), in contrast, makes it clear that the viewer did not find the video to be funny at all, garnering an `RNF` tag: a relevant (`R`) negative (`N`) comment about funniness (`F`). Perhaps a bit trickier, the comment in (11c) is `RNN`, being relevant (`R`) and expressive of a negative opinion (`N`), but not commenting on the funniness of the video (`N`). While sometimes it is challenging to sort out general from humorous sentiment, here the general negative opinion is obvious but the humor is not.

(11) a. The most hilarious video EVER!
b. did not laugh once. just awful stuff.
c. I DO NOT LIKE HOW THAY DID THAT

Turning to comments which do not use all levels of annotation: The comment in (12a) is a quote from the video and is tagged as being relevant and a quote: `RQ`. Finally, there are clear irrelevant cases, requiring only one level of annotation. The comment in (12b), for example, is tagged as `I`: It is not about the content of the video, and this annotation will not move on.

(12) a. MCOOOOYYY!!!
b. Subscribe to my channel

Difficult Cases Other comments prove to be more difficult to make judgments. One concept that underwent several iterations of fine-tuning was that of relevance. As discussed in section 3.1, certain aspects of a video are irrelevant, though determining which ones are or are not can be debatable.

Consider the discussion of actors, specifically as to whether a comment refers to the specific role in the video or the overall quality of their work. In (13a), for example, the comment is about the comedian as a person, not relevant to his performance in the video (`I`). We can see this more clearly in the distinction between the comment in (13b), which is `I`rrelevant, and the one in (13c), which is `R`elevant.

(13) a. Almost reminds me of Jim Carry ! :) lol she looks just like him Girl version
b. Kevin hart is always awesome !
c. I love kevin hart in this video

d. The name of this video is Ironic considering its coming from Cracked ;)

e. Homeschooling sucks worst mistake I ever made and I lost almost all social contact until college. It's great academically but terrible socially

From a different perspective, consider the video's title, e.g., in (13d): As we consider the title a part of the content of the video (in this case displaying an ironic situation), the comment is considered relevant. The emoticon suggests a positive emotion, and there is no reference to funniness (`RPN`).

Perhaps the most challenging conceptual issue with relevance is to distinguish the *topic* of the video from the *contents*. The comment in (13e) seems strongly negative, but, while it discusses the topic of the video (a comedy sketch on homeschooling), the opinion concerns the topic more generally, not the video itself: `I`.

Moving beyond relevance, other comments present challenges in teasing apart the opinion towards the video's contents versus opinions about other matters. The comment in (14a), for example, is relevant because it is about the video's content. This backhanded compliment must be counted as positive because, while insulting to all women, the user is complimenting the woman in the video. Thus, the comment is annotated `RPF`.

(14) a. Very funny for a woman

b. 104 people are [f***ing] dropped!

The comment in (14b) has a less direct interpretation, outside of the YouTube commenting context, in that it refers to the thumbs up/down counts for the video. While the comment does not directly address the content of the video, it indirectly does by referencing fraternity habits from the video (as in "dropped from pledging a fraternity"). The comment is annotated as positive despite the negative tone expressed because it implies that the 104 people who downvoted the video are wrong. Consequently, the comment is labeled `RPN`.

Once relevance and sentiment have been determined, there are still issues in terms of whether the comment is about funniness. In (15a), for instance, the relevant comment is clearly positive, but its status as being about funniness is unclear, necessitating the label `RPU`. Likewise, the comment in (15b) is negative with unclear funniness (`RNU`).

(15) a. Dude this go a three-pete! awesome!

b. Such a dated reference

c. Smiled...never laughed

From a different perspective, the comment in (15c) conveys a certain ambivalence, both about the sentiment (positive, but not overwhelmingly so) and about the funniness, distinguishing either between different kinds of humor or between humor/funniness and something else (i.e., something that induces smiling but not

laughing). In such cases, we annotate it as RPU, showing that the uncertain label helps deal with the gray area between clearly about funniness and clearly not.[2]

The comments in this section are cases that could only be annotated after an intense discussion of the annotation scheme and a clarification of the annotation guidelines. Additionally, the comments show that annotating for funniness or sentiment based on sentiment-bearing words is often not sufficient and can be misleading; (13e) is an example of this, as an irrelevant comment is filled with negative sentiment words. We need to consider the underlying intention of the comment, often expressed only implicitly. While this means an automatic approach to classifying the comments will be extremely challenging, such types of cases show that our current scheme is robust enough to handle these difficulties.

5 Annotator Differences and Machine Learning Quality

We have observed that, despite intensive exposure to the annotation scheme, our annotators make different decisions.[3] Consequently, we decided to investigate whether the differences in the annotations between annotators have an influence on automatic classification. If such decisions have an influence on the automatic annotations, we may need to adapt our guidelines further to increase agreement. If the differences between annotators do not have (much of) an effect on the automatic learner, we may be able to continue with the current annotations.

Since the goal is to investigate the influence of individual annotator decisions, we perform a tightly controlled experiment, using six videos[4] and four different annotators. To gauge the variance amongst different annotation styles, we performed machine learning experiments *within* each video and annotator. The task of the machine learner is to determine the complex label resulting from the concatenation of the three levels of annotation. Thus, we have four separate experiments, one for each annotator, and we compare results across those. This means the task is relatively easy because the training comments originate from the same video as the test comments, and out-of-vocabulary words are less of an issue, as well as video-specific funniness indicators being present (e.g., mentions of "twelve" in the Homeschooling video). But the task is also difficult because of the extremely small size of the training data given the fine granularity of the target categories. For each video and annotator, we perform threefold cross-validation.

We use Gradient Boosting Decision Trees as a classifier, as implemented in Scikit-learn [7] (http://scikit-learn.org/stable/) to predict the tripartite labels. We conduct experiments using default settings with no parameter optimiza-

[2]We only provide a single annotation for each comment, giving the most specific funniness level appropriate for any sentence in the comment; while future work could utilize comments on a sentential or clausal level, we did not encounter many problematic cases.

[3]Space precludes a full discussion of inter-annotator agreement (IAA); depending upon how one averages IAA scores across four annotators, one finds Level 1 agreement of 83% and agreement for all three levels of annotation around 61–62%.

[4]IDs: V97lvUKYisA, Q9UDVyUzJ1g, zvLpXIYLrec, cFkIJBVZ4_w, WmIS_icNcLk, rLw-9dpHtcU

Video	Size	A1	A2	A3	A4	Avg.
Tig Notaro	67	31.81	42.42	37.88	34.85	36.74
J. Phoenix	99	42.86	43.88	50.00	38.78	43.88
Water	100	34.34	36.36	45.45	35.35	37.87
Homeschooling	87	51.16	56.98	50.00	51.16	52.33
Kevin Hart	101	33.33	32.00	41.00	37.00	35.83
Spider	100	55.56	47.47	40.40	46.46	47.47

Table 2: Classification results (%) per video and annotator and on average (using default parameters)

tion. This setting is intended to keep the parameters stable across all videos to allow better comparability. We use bag-of-words features (recording word presence/absence) since they usually establish a good baseline for machine learning.

Results Table 2 shows the results for the experiments using default settings for the classifier. Comparing across videos, *Homeschooling* has the highest average machine learner performance while *Tig Notaro* and *Kevin Hart* have the lowest.[5]

Comparing across videos, the results show that all videos seem to present a similar level of difficulty, with *Homeschooling* being the easiest video and *Kevin Hart* the most difficult, based on averaged classification results. However, accuracies vary considerably between annotators. For example, A1's annotations for *Tig Notaro* resulted in dramatically lower ML accuracies than A2's. For *Spider*, the opposite is true. The differences between the highest and lowest result per video can be as high as 15% (absolute), for the *Spider* video. These results make it clear that the different choices of annotators have a considerable effect on the accuracy of the machine learner.

6 Conclusion

In this paper, we have presented a tripartite annotation scheme for annotating comments about the funniness of YouTube videos. We have shown that humor is a complex concept and that it is necessary to annotate it in the context of relevance to the video and of sentiment. Our investigations show that differences in annotation have a considerable influence on the quality of automatic annotations via a machine learner. This means that reaching consistent annotations between annotators is of extreme importance.

[5]If one compares the average results to IAA rates, there is no clear correlation, indicating that IAA is not the only factor determining classifier accuracy.

References

[1] Salvatore Attardo. Semantics and pragmatics of humor. *Language and Linguistic Compass*, 2(6):1203—1215, 2008.

[2] Diana Inkpen and Carlo Strapparava, editors. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, Los Angeles, CA, June 2010.

[3] Bing Liu. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

[4] Rada Mihalcea and Stephen Pulman. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, Mexico City, 2007. Springer.

[5] Subhabrata Mukherjee and Pushpak Bhattacharyya. YouCat: Weakly supervised youtube video categorization system from meta data & user comments using Wordnet & Wikipedia. In *Proceedings of COLING 2012*, pages 1865–1882, Mumbai, India, December 2012.

[6] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc., Sebastopol, CA, 2013.

[9] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. How useful are your comments? - analyzing and predicting youtube comments and comment ratings. In *Proceedings of WWW 2010*, Raleigh, NC, 2010.

[10] Rachele Sprugnoli, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 2015.

[11] Jeffrey Charles Witt. The sentences commentary text archive: Laying the foundation for the analysis, use, and reuse of a tradition. *Digital Humanities Quarterly*, 10(1), 2016.