# Humor in Collective Discourse: Unsupervised Funniness Detection in the New Yorker Cartoon Caption Contest

**Dragomir Radev[1], Amanda Stent[2], Joel Tetreault[2], Aasish Pappu[2]**
**Aikaterini Iliakopoulou[3], Agustin Chanfreau[3], Paloma de Juan[2], Jordi Vallmitjana[2]**
**Alejandro Jaimes[2], Rahul Jha[1], Robert Mankoff[4]**

[1] University of Michigan, [2] Yahoo! Research, [3] Columbia University, [4] The New Yorker
radev@umich.edu, {stent,tetreaul,aasishkp}@yahoo-inc.com, {ai2315,ac3680}@columbia.edu
{pdejuan,jvallmi,ajaimes}@yahoo-inc.com, rahuljha@umich.edu, bob_mankoff@newyorker.com

**Abstract**

The New Yorker publishes a weekly captionless cartoon. More than 5,000 readers submit captions for it. The editors select three of them and ask the readers to pick the funniest one. We describe an experiment that compares a dozen automatic methods for selecting the funniest caption. We show that negative sentiment, human-centeredness, and lexical centrality most strongly match the funniest captions, followed by positive sentiment. These results are useful for understanding humor and also in the design of more engaging conversational agents in text and multimodal (vision+text) systems. As part of this work, a large set of cartoons and captions is being made available to the community.

**Keywords:** collective discourse, computational humor, creativity

## 1. Introduction

The New Yorker Cartoon Caption Contest has been running for more than 10 years. Each week, the editors post a cartoon (cf. Figures 1 and 2) and ask readers to come up with a funny caption for it. They pick the top 3 submitted captions and ask the readers to pick the weekly winner. The contest has become a cultural phenomenon and has generated a lot of discussion as to what makes a cartoon funny (at least, to the readers of the New Yorker). In this paper, we take a computational approach to studying the contest to gain insights into what differentiates funny captions from the rest. We developed a set of unsupervised methods for ranking captions based on features such as originality, centrality, sentiment, concreteness, grammaticality, human-centeredness, etc. We used each of these methods to independently rank all captions from our corpus and selected the top captions for each method. Then, we performed Amazon Mechanical Turk experiments in which we asked Turkers to judge which of the selected captions is funnier.

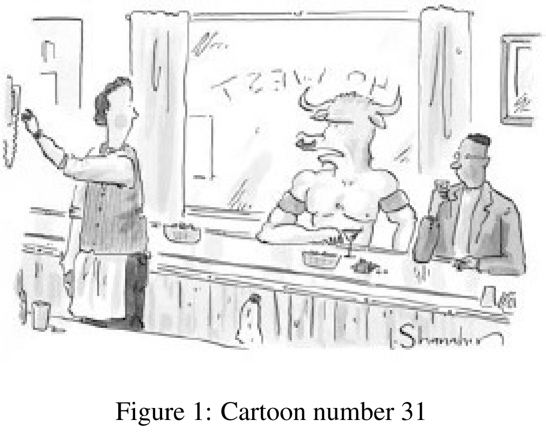


Figure 1: Cartoon number 31

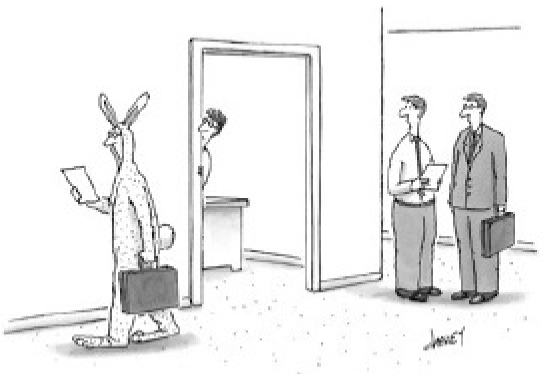


Figure 2: Cartoon number 32

## 2. Related Work

In early work, Mihalcea and Strapparava (2005) investigate whether classification techniques can distinguish between humorous and non-humorous text. Training data consisted of humorous one-liners (15 words or less), and non-humorous one-liners, which are derived from Reuters news titles, proverbs, and sentences from the British National Corpus. They looked at features such as alliteration, antonymy and adult slang.

Mihalcea and Pullman (2007) took this work further. They looked at four semantic classes relevant to human-centeredness: persons, social groups, social relationships, and personal pronouns. They showed that social relationships and personal pronouns have high prevalence in humor. Mihalcea and Pullman also looked at sentiment; they found that humor tends to have a strong negative orientation (especially in the case of long satirical text, but regular text also shows some tendency toward the negative). Reyes et al. (2009) used these same features as well as others to build a humor taxonomy.

Raz (2012) classified tweets by type and topic, while Barbieri and Saggion (2014) focused on classifying tweets into Irony, Education, Humour, and Politics. Zhang and Liu (2014), also looking at tweets, used a set of manually crafted features based on influential humor theories, linguistic norms, and affective dimensions.

In a more recent paper, Shahaf et al. (2015) describe work conducted in parallel with ours. They also created a corpus of pairs of captions (one funny and one not funny) and determined that the funny captions are statistically different from the unfunny ones in the pairs. Then they built a classifier that picks the funnier one in the pair at a 69% accuracy.

Our work differs from previous research in several ways. First, most previous work has focused on automatically distinguishing between humorous and non-humorous text. In our case, the goal is to rank humorous texts (and assess *why* they are funny), not perform binary classification. Second, we're not aware of any work that deals specifically with cartoon captions, and although our methods are not specific to captions, we include features based on the objects depicted in the cartoons.

## 3. Data

We have access to a corpus of more than 2M captions for more than 400 contests run since 2005. For our experiments we picked a subset of 50 cartoons and 298,224 captions. Our data set includes, for each contest, the following:

- the cartoon itself
- 5,000+ captions, tokenized using ClearNLP 2.0 (Choi and Palmer, 2012)
- the three selected captions, including the winning caption
- the most frequent n-grams in the captions
- manually labeled objects that are visible in the cartoon
- tfidf scores for all captions
- "antijokes" from two sites (AlInLa[1] and Radosh[2]), devoted to "unfunny" captions

## 4. Experimental Setup

We developed more than a dozen unsupervised methods for ranking the submissions for a given contest. As controls, we use the three captions selected by the editors of the New Yorker as well as antijokes. For all methods, we broke ties randomly. Some of our methods can be used in two different directions (e.g., `CU2` favors the most positive captions whereas `CU2R` the most negative ones). The methods and baselines are split into five groups: OR=originality based, GE=generic, CU=content, NY=original New Yorker contest, CO=control.

- (`OR1 & OR1R`) similarity to contest centroid
- (`OR2 & OR2R`) highest/lowest lexrank
- (`OR3 & OR3R`) largest/smallest cluster
- (`OR4`) highest average tfidf
- (`CU1`) presence of Freebase entities (Bollacker et al., 2008)
- (`CU2 & CU2R`) caption sentiment
- (`CU3`) human-centeredness
- (`GE1`) most syntactically complex
- (`GE2`) most concrete (i.e., refers to objects present in the cartoon)
- (`GE3 & GE3R`) unusually formatted text
- (`NY1`) first place official
- (`NY2`) second place official
- (`NY3`) third place official
- (`CO2`) antijokes

### 4.1. Originality-based methods

We built a lexical network out of the captions for each contest. We used LexRank (Erkan and Radev, 2004) to identify the most central caption in each contest (method `OR1`) and the one with the highest lexrank score (method `OR2`). We also used Louvain, a graph clustering method (Blondel et al., 2008), previously used in King et al. (2013), to cluster the captions in each contest thematically; the sizes of these clusters comprise method `OR3`. The tfidf scores used to build the lexical network are used in method `OR4`.

Figure 4 shows the pairwise similarities for the captions in the mini-corpus. The seven clusters are identified by the Louvain method. Solid lines represent high cosine similarity between a pair of captions.

The captions in the mini-corpus are shown in Figure 3. The seven clusters in Figure 5 are identified by the Louvain method. Solid lines represent high cosine similarity between a pair of captions.

### 4.2. Content-based methods

For `CU1`, we annotated the captions for Freebase entities by querying noun-phrases (within a caption) over Freebase indexed entities. We scored each caption using idf $*$ Freebase score, where the Freebase score captures relevance.

To compute the sentiment polarity of each caption (method `CU2`), we used Stanford CoreNLP (Manning et al., 2014) to annotate each sentence with its sentiment from 0 (very negative) to 4 (very positive). Only 13.20% had positive polarity; 51.09% had negative polarity, and the rest were neutral.

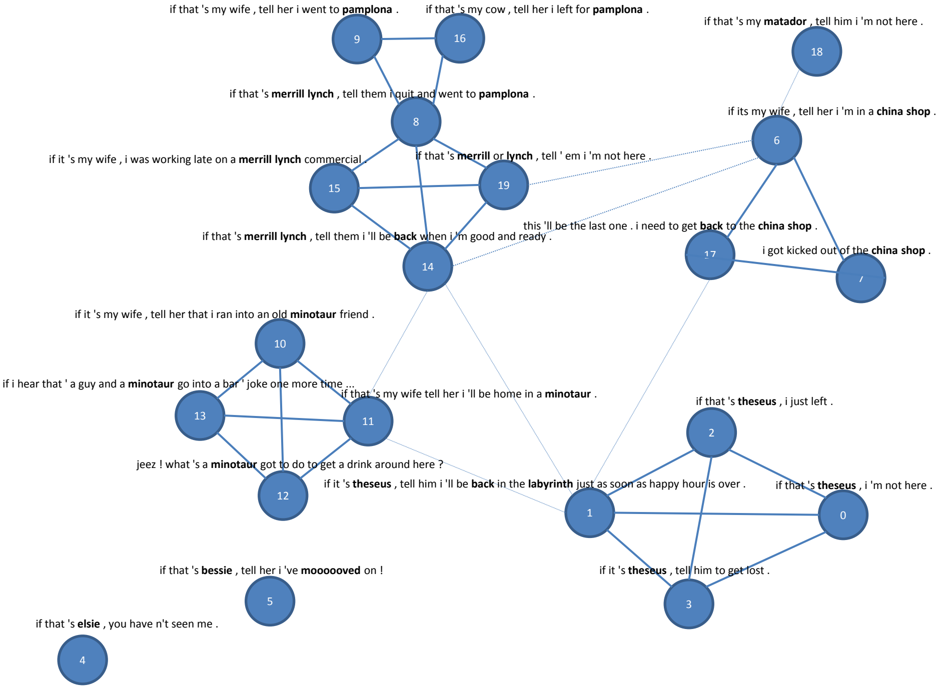
For human-centeredness (method `CU3`), we followed the method described in Mihalcea and Pullman (2007). We used WordNet (Miller, 1995) to list all the word forms derived from the $\{$*person, individual, someone, somebody, mortal, human, soul*$\}$ synset ("people" set), as well as those belonging to the $\{$*relative, relation*$\}$ synset ("relatives" set).

[1] http://alinla.blogspot.com/

[2] http://www.radosh.net/

```
 0 0    if that ’s theseus , i ’m not here .
 1 0    if it ’s theseus , tell him i ’ll be back in the labyrinth just as soon as happy hour is over .
 2 0    if that ’s theseus , i just left .
 3 0    if it ’s theseus , tell him to get lost .
 4 1    if that ’s elsie , you have n’t seen me .
 5 2    if that ’s bessie , tell her i ’ve moooooved on !
 6 3    if its my wife , tell her i ’m in a china shop .
 7 3    i got kicked out of the china shop .
 8 5    if that ’s merrill lynch , tell them i quit and went to pamplona .
 9 5    if that ’s my wife , tell her i went to pamplona .
10 4    if it ’s my wife , tell her that i ran into an old minotaur friend .
11 4    if that ’s my wife tell her i ’ll be home in a minotaur .
12 4    jeez ! what ’s a minotaur got to do to get a drink around here ?
13 4    if i hear that ’ a guy and a minotaur go into a bar ’ joke one more time ...
14 5    if that ’s merrill lynch , tell them i ’ll be back when i ’m good and ready .
15 5    if it ’s my wife , i was working late on a merrill-lynch commercial .
16 5    if that ’s my cow , tell her i left for pamplona .
17 3    this ’ll be the last one . i need to get back to the china shop .
18 6    if that ’s my matador , tell him i ’m not here .
19 5    if that ’s merrill or lynch , tell ’ em i ’m not here .
```

Figure 3: Subset of the captions for contest number 31, labeled by thematical cluster (column 2). 0 - theseus, 1 - elsie, 2 - bessie, 3 - china shop, 4 - minotaur, 5 - merrill lynch, 6 - matador.



if that 's my wife , tell her i went to **pamplona** .
9
if that 's my cow , tell her i left for **pamplona** .
16
if that 's my **matador** , tell him i 'm not here .
18
if that 's **merrill lynch** , tell them i quit and went to **pamplona** .
8
if its my wife , tell her i 'm in a **china shop** .
6
if it 's my wife , i was working late on a **merrill lynch** commercial .
15
if that 's **merrill or lynch** , tell ' em i 'm not here .
19
this 'll be the last one . i need to get **back** to the **china shop** .
if that 's **merrill lynch** , tell them i 'll be **back** when i 'm good and ready .
14
17
i got kicked out of the **china shop** .
7
if it 's my wife , tell her that i ran into an old **minotaur** friend .
10
if i hear that ' a guy and a **minotaur** go into a bar ' joke one more time ...
if that 's my wife tell her i 'll be home in a **minotaur** .
13
11
if that 's **theseus** , i just left .
2
jeez ! what 's a **minotaur** got to do to get a drink around here ?
if it 's **theseus** , tell him i 'll be **back** in the **labyrinth** just as soon as happy hour is over .
if that 's **theseus** , i 'm not here .
12
1
0
if that 's **bessie** , tell her i 've **moooooved** on !
if it 's **theseus** , tell him to get lost .
5
3
if that 's **elsie** , you have n't seen me .
4

Figure 5: Lexical network for contest 31.

We excluded personal pronouns, as 75.96% of the captions contained at least one. We also accounted for any proper names as part of the “people” set. 25.33% of the captions mentioned at least one “person”, but only 3.60% contained a word from the “relatives” set.

## 4.3. Generic methods

We computed syntactic complexity (GE1) using (Charniak and Johnson, 2005). For concreteness (GE2), two of the authors of this paper labeled all the objects in each of the 50 cartoons used in our evaluation. We then computed how often any of those objects were referred to (with a nominal NP) in each caption. We computed GE3 by counting punctuation marks and unusually formatted (e.g. very long) words in each caption.

## 5. Evaluation

We used Amazon Mechanical Turk (AMT) to compare the outputs of the different methods and the baselines. Each AMT HIT consisted of one cartoon as well as two captions, A and B (produced by one of the 18 methods and baselines). The turkers had to determine which of the two captions is funnier. They were given four options - “A is funnier”, “B is funnier”, “both are funny”, “neither is funny”. They did not know which method was used to produce caption A or B. All pairs of captions from our methods were compared for each cartoon, and each HIT (pair) was assessed by 7 Turkers.

We report on three evaluations in Table 1. Each evaluation ($n_i$, $s_i$ pair) corresponds to the number of votes in favor of the given method minus the number of votes against. So the first set corresponds to pairs in which, out of seven

| Category | Code | Method | $n_4$ | $s_4$ | $n_3$ | $s_3$ | $n$ | $s$ |
|---|---|---|---|---|---|---|---|---|
| Centrality | OR1R | least similar to centroid | 308 | -2.73 | 453 | -2.14 | 846 | -1.26 |
| | **OR2** | highest lexrank | 302 | **1.39** | 457 | **1.11** | 846 | **0.59** |
| | OR2R | smallest lexrank | 317 | -0.61 | 450 | -0.58 | 846 | -0.29 |
| | OR3R | small cluster | 468 | -4.40 | 581 | -3.94 | 848 | -2.85 |
| | OR4 | tfidf | 474 | -4.93 | 596 | -4.36 | 850 | -3.24 |
| New Yorker | **NY1** | official winner | 314 | **3.57** | 466 | **2.96** | 847 | **1.78** |
| | **NY2** | official runner up | 330 | **3.24** | 463 | **2.60** | 845 | **1.54** |
| | **NY3** | official third place | 276 | **2.29** | 435 | **1.57** | 842 | **0.89** |
| General | GE1 | syntactically complex | 268 | -0.10 | 406 | -0.14 | 846 | -0.70 |
| | GE2 | concrete | 259 | -0.33 | 427 | -0.41 | 844 | -0.26 |
| | GE3R | well formatted | 296 | 0.81 | 446 | 0.61 | 846 | 0.31 |
| Content | CU1 | freebase | 290 | 0.26 | 424 | 0.17 | 840 | 0.07 |
| | **CU2** | positive sentiment | 268 | **1.21** | 396 | **0.83** | 836 | **0.46** |
| | **CU2R** | negative sentiment | 298 | **1.69** | 445 | **1.30** | 826 | **0.70** |
| | **CU3** | people | 276 | **1.45** | 409 | **1.24** | 834 | **0.68** |
| Control | CO2 | antijoke | 259 | 0.27 | 394 | -0.04 | 822 | -0.09 |

Table 1: Comparison between the methods. Score $s_4$ corresponds to pairs for which the seven judges agreed more significantly (a difference of 4+ votes). Score $s_3$ requires a difference of 3+ votes. Score $s$ includes all pairs (about 850 per method, minus a small number of errors). The best methods (CU2R, CU3, OR2, and CU2) are in bold.

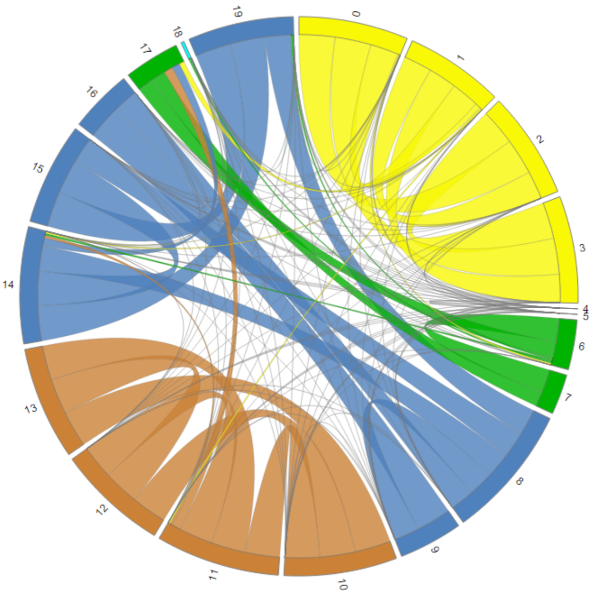


Figure 4: Clustering of the mini corpus

judges, there was a difference of at least 4 votes in favor of one or the other caption. This level of significant agreement happened in 5,594/15,154 cases (36.9% of the time). A difference of at least 3 votes happened in 8,131/15,154 pairs (53.6%). The third evaluation corresponds to all pairwise comparisons, including ties. $n_i$ refers to the number of times the above constraint for $i$ is met and score $s_i$ is calculated by averaging the number of votes in favor minus the number of votes against for each $n_i$. The probability that a random process will generate a difference of at least 4 votes (excluding ties) is 12.5%.

## 6. Conclusion and Data Release

We compared over a dozen methods for selecting the funniest caption among 5,000 submissions to the New Yorker caption contest. Using side by side funniness assessments from AMT, we found that the methods that consistently select funnier captions are negative sentiment, human-centeredness, and lexical centrality. Not surprisingly, knowing the traditions of the New Yorker cartoons, negative captions were funnier than positive captions. Captions that relate to people were consistently deemed funnier. The first two methods (negative sentiment and human-centeredness) are consistent with the findings in Mihalcea and Pullman (2007). More interestingly, we also showed that captions that reflect the collective wisdom of the contest participants outperformed semantic outliers. The next two strongest features were positive sentiment and proper formatting.

We are making our corpus public for research and for a shared task on funniness detection. The corpus includes our 50 selected cartoons, more than 5,000 captions per cartoon, manual annotations of the entities in the cartoons, automatically extracted topics from each contest, and the funniness scores.

## 7. Future Work

In this paper, we used unsupervised methods for funniness detection. We will next explore supervised and ensemble methods. (However, ensemble methods may not work for this task as captions may be funny in different ways; for example, of two equally funny captions, one may be funny-absurd and the other funny-ironic.) We will also explore pun recognition (e.g., “Tell my wife I’ll be home in a *minotaur.*”), other creative uses of language, as well as more semantic features.

## 8. Bibliographical References

Barbieri, F. and Saggion, H. (2014). Automatic detection of irony and humour in twitter. In *Proceedings of the International Conference on Computational Creativity*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of SIGMOD*.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL*.

Choi, J. D. and Palmer, M. (2012). Fast and robust part-of-speech tagging using dynamic model selection. In *Proceedings of ACL*.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

King, B., Jha, R., Radev, D. R., and Mankoff, R. (2013). Random walk factoid annotation for collective discourse. In *Proceedings of ACL*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL*, pages 55–60.

Mihalcea, R. and Pulman, S. G. (2007). Characterizing humour: An exploration of features in humorous texts. In *Proceedings of CICLing*.

Mihalcea, R. and Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of HLT/EMNLP*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, Nov.

Raz, Y. (2012). Automatic humor classification on Twitter. In *Proceedings of NAACL/HLT*.

Reyes, A., Rosso, P., and Buscaldi, D. (2009). Evaluating humorous features: Towards a humour taxonomy. In *Proceedings of the Indian International Conference on Artificial Intelligence*.

Shahaf, D., Horvitz, E., and Mankoff, R. (2015). Inside jokes: Identifying humorous cartoon captions. In *Proceedings of SIGKDD*.

Zhang, R. and Liu, N. (2014). Recognizing humor on twitter. In *Proceedings of ACM CIKM*.