Evaluating Humor Features on Web Comments

Antonio Reyes[1], Martin Potthast[2], Paolo Rosso[1], Benno Stein[2]

[1]Natural Language Engineering Lab — ELiRF
Universidad Politécnica de Valencia, Spain
{areyes,prosso}@dsic.upv.es

[2]Web Technology & Information Systems
Bauhaus-Universität Weimar, Germany
{martin.potthast,benno.stein}@uni-weimar.de

Abstract

Research on automatic humor recognition has developed several features which discriminate funny text from ordinary text. The features have been demonstrated to work well when classifying the funniness of single sentences up to entire blogs. In this paper we focus on evaluating a set of the best humor features reported in the literature over a corpus retrieved from the Slashdot Web site. The corpus is categorized in a community-driven process according to the following tags: funny, informative, insightful, offtopic, flamebait, interesting and troll. These kinds of comments can be found on almost every large Web site; therefore, they impose a new challenge to humor retrieval since they come along with unique characteristics compared to other text types. If funny comments were retrieved accurately, they would be of a great entertainment value for the visitors of a given Web page. Our objective, thus, is to distinguish between an implicit funny comment from a not funny one. Our experiments are preliminary but nonetheless large-scale: 600,000 Web comments. We evaluate the classification accuracy of naive Bayes classifiers, decision trees, and support vector machines. The results suggested interesting findings.

1. Introduction

Today, the Web is the major source of data for many scientific and non-scientific areas: blogs, bulletin boards, wikis, social networks, and the like are rich resources for topic-centric but also for non-topic-driven retrieval research. With respect to the latter, e.g. the research of (Pang et al., 2002) shows the importance of movie reviews for sentiment analysis, and (Balog et al., 2006) demonstrate how to exploit user-generated tags on blogs to analyze irregularities in the moods of bloggers.

Our paper focuses on the retrieval of humorous texts—more precisely, on the retrieval of funny comments on Web items. Comments can be found on almost every large Web site; they impose a new challenge to humor retrieval since they come along with unique characteristics compared to other text types. If funny comments were retrieved accurately, they would be of a great entertainment value for the visitors of a given Web page. To this end, we introduce a new large-scale corpus for humor retrieval: the Slashdot news Web site which contains human-annotated funny comments on a large scale.

The following sections review related work (Section 2.), introduce the used text features (Section 3.), report on our experiments and the achieved results (Section 4.), and discuss the findings (Section 5.).

2. Related Work

Humor retrieval research pursues tow research goals: (i) the automatic generation of humorous contents (Binsted and Ritchie, 1997; Stock and Strapparava, 2005) and (ii) the automatic recognition of humor (Mihalcea and Strapparava, 2006a; Mihalcea and Pulman, 2007).

With respect to the latter, a number of features have been proposed which discriminate between funny and ordinary texts. Mihalcea and Strapparava (2006a) use the appearance of alliterations, antonyms, and sexual content to distinguish one-liners from proverbs, news titles, and sentences from both the British National Corpus and the Open Mind Common Sense corpus. Mihalcea and Pulman (2007) evaluate how human-centric vocabulary and negative polarity affect the classification accuracy when discriminating one-liners and humorous news articles from serious texts. Reyes et al. (2009a) evaluate semantic ambiguity and affective information in order to classify blogs with respect to the bloggers' moods. Other researchers evaluate text similarity, writing style, and idiomatic expressions (Sjöbergh and Araki, 2007), text length, n-gram representations, and bag-of-words representations (Buscaldi and Rosso, 2007), as well as keyness and discriminative items (Reyes et al., 2009b).

3. Humor Model and Evaluation Corpus

In our humor model we employ a selection of the best-performing humor features found in the literature, along with new features that are unique for comment text. These new features are terms which are used in natural language to express certain kinds of feelings; the terms divide into the following five categories:

1. sexual terms from the sexuality domain (Bentivogli et al., 2004);
2. terms with negative polarity (Esuli and Sebastiani, 2006);
3. semantic ambiguous terms, based on sense dispersion (Reyes et al., 2009b);

4. terms that reflect emotions, based on the affective term categories (Strapparava and Valitutti, 2004);
5. slang and emoticons, e.g., expressions like “LOL” or “`:-)`”.

In an offline pre-processing step the terms that belong to these categories are filtered, based on the currently most representative evaluation corpus in humor recognition, the one-liners corpus (Mihalcea and Strapparava, 2006a).[1] If a term occurs less than 50 times in this corpus it is discarded from the vocabulary. Given the pre-processed vocabulary, every comment is represented as a frequency-weighted term vector. The underlying hypothesis is that those features which best indicate humor for one-liners will also be useful for comments.

3.1. Evaluation Corpus

Our evaluation corpus consists of about 3.8 million comments retrieved from the Slashdot news Web site. It includes all comments on articles published between January 2006 and June 2008. Comments on Slashdot are categorized in a community-driven process. The comment categories include the following tags: funny, informative, insightful, interesting, off-topic, flamebait, and troll.[2]
The following comments are concrete examples about how the Slashdot community, depending on the meaning they want to communicate, categorize their own comments by means of the previous tags.

- *Re:Number of movies (Score:5, Insightful).*

  “I believe that prior to this particular month, HD-DVD was consistently ahead of Blu-Ray. Declaring a winner based on a single months’ worth of statistics (especially at this early point when both formats are in their infancy) is utterly idiotic.”

- *Re:Number of movies (Score:1, Interesting).*

  “True. However, it can be used as a tool to gage the trend to try to predict WHERE the winning format will fall.”

- *Re:Number of movies (Score:2, Funny).*

  “So let me get this straight: A single data point can be used as a "tool" to gage the trend? No shit?”

- *Re:Number of movies (Score:2, Funny).*

  “6 months of data is a single data point? No shit? It’s not a single data point. It’s the volume of title sales over 6 months. RTFA and maybe... just MAYBE click the links.”

The amount of comments on Slashdot does not allow for every comment to be categorized, so that we restrict ourselves to the 1.068,953 categorized comments. They are divided into four classes: funny, informative, insightful, and negative. The latter contains comments from categories off-topic, flamebait, interesting and troll. The funny class is the smallest of the four; it contains 159,153 comments. In order to avoid problems related to class imbalance, samples of 150,000 comments from each of the other three classes are employed in the experiments, i.e., 600,000 comments in total. Figure 1 depicts the representativeness of the set of features regarding the four classes.

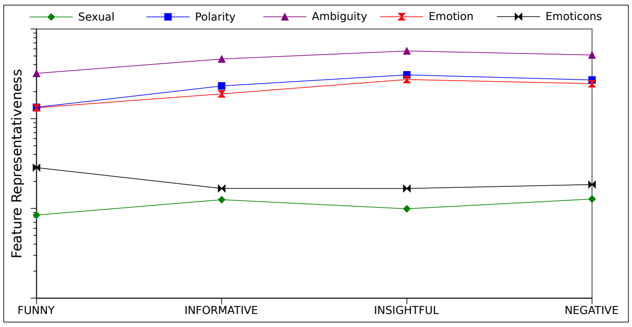

Figure 1: Feature representativeness per class.

4. Experiments and Results

The experiments are carried out with three classifier technologies: naive Bayes, decision trees, and support vector machines (SVM). The training sets contain 100,000 comments per class, the test sets contain 50,000 comments per class. Each classifier is evaluated using different sets of features. The following schema summarizes the features and the order in which they are assessed:

s_1 sexual-content and semantic ambiguity

s_2 sexual-content, semantic ambiguity, and polarity

s_3 sexual-content, semantic ambiguity, polarity, and emotions

s_4 all features

All classifications experiments consider the classes funny versus informative, insightful, and negative respectively. The Tables 1-3 comprise the results.
From the results it can be inferred that the features discriminate less well compared to the classification setting where one-liners and news titles are being told apart. Note, however, that the most similar classes are funny and informative, whereas the negative class and the insightful class are more different. On the other hand, it is interesting to notice that, despite the results reported in the literature, the “emotions” feature does not improve the classification accuracy over these classes, whereas the new features “slang” and “emoticons” improve classification accuracy. Also note that this feature is less representative than the features “ambiguity” and “polarity” (cf. Figure 1).

5. Discussion

Our a-priori intuition is to transfer well-known humor features to evaluate their discriminative power in distinguishing funny comments from ordinary ones. The results, however, show that these features, despite their good performance on one-liners, are not very useful for comments. We

[1] Due to the lack of a gold standard in computational humor recognition, we decided to use this corpus, given the excellent results reported on it in the literature.

[2] This corpus has firstly been used for measuring the descriptiveness of Web comments (Potthast, 2009).

Table 1: Classification accuracy of funny vs. informative.

Exp.	Bayes	SVM	REPTree
s_1	57.15%	57.16%	57.16%
s_2	57.35%	57.38%	57.36%
s_3	58.03%	57.38%	57.29%
s_4	58.26%	57.94%	58.31%

Table 2: Classification accuracy of funny vs. insightful.

Exp.	Bayes	SVM	REPTree
s_1	62.19%	62.25%	62.25%
s_2	62.66%	62.43%	62.74%
s_3	62.39%	62.52%	62.94%
s_4	63.08%	62.97%	63.52%

Table 3: Classification accuracy of funny vs. negative.

Exp.	Bayes	SVM	REPTree
s_1	60.37%	60.36%	60.37%
s_2	60.54%	60.41%	60.54%
s_3	60.13%	60.37%	60.54%
s_4	60.48%	60.89%	61.33%

explain this behavior by two correlated reasons: (i) the negative data sets and, (ii) the kind of linguistic strategies profiled: one-liners on the one hand versus comments on the other hand.

Regarding the first reason, observe that the best results reported in the literature have been achieved on data sets from completely different sources, i.e., one-liners versus news titles or sentences from the British National Corpus. These are data sets with similar structures, but also with significant differences regarding topic, vocabulary, or target audience. In our case, the not-funny training examples are of the same text type as the funny ones. They hence share a common source, namely the Slashdot corpus, whereas the only difference are users tags. Altogether the examples share more common aspects than differences.

Regarding the second reason, consider that one-liners and funny comments focus on two different linguistic strategies to achieve their effect. Both imply an underlying funny sense, but the way humor is produced is different. Humor in one-liners is caused by linguistic strategies such as ambiguity, irony, sarcasm, apart from cultural and social information. Humor in comments is introduced with a response to a comment of someone else; the underlying mechanism that introduces humor relies on making clear a discrepancy between two particular points of view. For instance, the sexual-content feature, which is relevant when classifying one-liners, is the least representative one in our classes, whereas emoticons, i.e., visual elements which imply the funny sense, are used rather often in funny comments (cf. Figure 1).

On the basis of these insights we decided to carry out another, straightforward experiment: 20 comment threads are randomly selected, each containing at least 30 funny comments, and the dispersion among the senses profiled by every thread is measured. We apply the following formula (Reyes et al., 2009b) to quantify the total sense dispersion per thread:

$$\delta(w_s) = \frac{1}{P(|S|,2)} \sum_{s_i,s_j \in S} d(s_i,s_j), \qquad (1)$$

where S is the set of synsets, $s_1, ..., s_n$, for word w; $P(n,k)$ is the number of permutations of n objects in k slots, and $d(s_i,s_j)$ is the length of the hypernym path between synsets (s_i,s_j). This measure quantifies the differences among the senses of a word considering the hypernym distance of the WordNet synsets. It relies on the hypothesis that a word with senses that differ significantly is more likely to be used to trigger metalinguistic information than a word with senses that differ slightly.

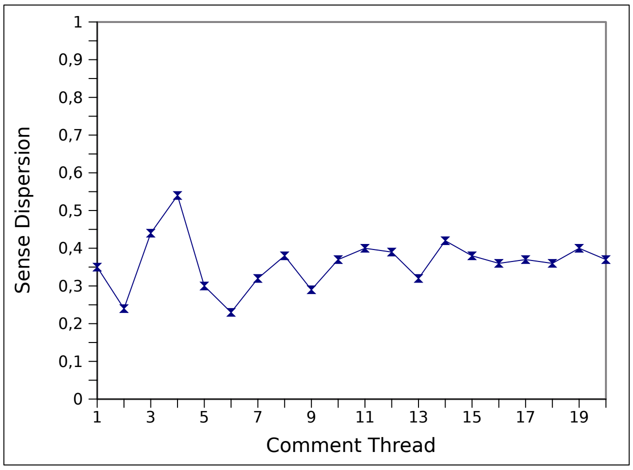

Figure 2: Sense Dispersion considering all the words in the comment thread.

The results in Figure 2 indicate a low dispersion among the senses of each comment thread, which means that the comments share more similarities than differences. For instance, except for one pair of threads, in the rest the sense dispersion barely exceeds 0.4: with increasing sense dispersion the divergences in the document increase as well. This observation supports the second reason about the low accuracy reached in our classifications experiments. Regarding the first reason, three classifiers (naive Bayes, decision tree, SVM) are trained considering 10,000 reviews extracted from the TripAdvisor data set (Baccianella et al., 2009), and 10,000 randomly selected funny comments. Each classifier is evaluated using the set which includes all the features (s_4). The attribute selection and principal components filters (Witten and Frank, 2005) are employed as well as the ten-fold cross validation method. Table 4 summarizes the results.

Although the number of documents classified is reduced, the results indicate that the consideration of a different negative data set improves the accuracy significantly.

6. Conclusions and Future Work

This paper evaluates the performance of the most discriminative features described in the research on automatic hu-

Table 4: Classification accuracy of hotel reviews vs. funny comments.

Exp.	Bayes	SVM	REPTree
s_4	73.43%	74.06%	73.17%

mor recognition in the field of Web comments. We distinguish between four classes of comments, using a set of five feature categories. The results show that the features have a limited performance in distinguishing funny comments from informative, insightful, and negative comments. We explain this with the negative data sets and the linguistic strategies employed between the "gold standard" and our positive set of funny comments. Our current work deals with the fact that a funny comment is often an answer either to the commented item or to another comment. Moreover, we investigate new features, such as those used for vandalism detection on Wikipedia.

Acknowledgments

This work is partially funded by the TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) project.

7. References

S. Baccianella, A. Esuli, and F. Sebastiani. 2009. Multifacet rating of product reviews. In *Proceedings of the 31st European Conference on Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 461–472. Springer.

K. Balog, G. Mishne, and M. Rijke. 2006. Why are they excited? identifying and explaining spikes in blog mood levels. In *European Chapter of the Association of Computational Linguistics (EACL 2006)*.

L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, pages 94–101. COLING.

K. Binsted and G. Ritchie. 1997. Computational rules for punning riddles. *Humour*, 10:25–75.

D. Buscaldi and P. Rosso. 2007. Some experiments in humour recognition using the italian wikiquote collection. In *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, volume 4578, pages 464–468.

A. Esuli and F. Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422.

R. Mihalcea and S. Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *8th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2007*, volume 4394, pages 337–347.

R. Mihalcea and C. Strapparava. 2006a. Technologies that make you smile: Adding humour to text-based applications. *IEEE Intelligent Systems*, 21(5):33–39.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

M. Potthast. 2009. Measuring the descriptiveness of web comments. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 724–725.

A. Reyes, P. Rosso, and D. Buscaldi. 2009a. Affect-based features for humour recognition. In *Proceedings of the 7th International Conference on Natural Language Processing ICON-09, (to be published)*.

A. Reyes, P. Rosso, and D. Buscaldi. 2009b. Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, 18(4).

J. Sjöbergh and K. Araki. 2007. Recognizing humor without recognizing meaning. In *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, volume 4578, pages 469–476.

O. Stock and C. Strapparava. 2005. Hahacronym: A computational humor system. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 113–116.

C. Strapparava and A. Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, volume 4, pages 1083–1086.

I. Witten and E. Frank. 2005. *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers. Elsevier.