# UC Merced

**Title**
Computational Laughing: Automatic Recognition of Humorous One-Liners

**Permalink**
https://escholarship.org/uc/item/6wg3b07c

**Journal**

**ISSN**

**Authors**
Miwa, Kazuhisa
Nakazawa, Daisuke
Saito, Hitomi

**Publication Date**
2005

Peer reviewed

# Computational Laughing:
# Automatic Recognition of Humorous One-liners

**Rada Mihalcea (rada@cs.unt.edu)**
Department of Computer Science, University of North Texas
Denton, Texas, USA

**Carlo Strapparava (strappa@itc.it)**
ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica
Povo, Trento, Italy

## Abstract

Humor is one of the most interesting and puzzling aspects of human behavior. Despite the attention it has received in fields such as philosophy, linguistics, and psychology, there have been only few attempts to create computational models for humor recognition or generation. In this paper, we bring empirical evidence that computational approaches can be successfully applied to the task of humor recognition. Through experiments performed on very large data sets, we show that automatic classification techniques can be effectively used to distinguish between humorous and non-humorous texts, with significant improvements observed over apriori known baselines.

## Introduction

Humor is an essential element in personal communication. Although strictly related to themes such as entertainment, fun, and emotion, it is an integral part of our lives, and arguably humans could not survive without it. Indeed, while it is merely considered a way to induce amusement, humor also has a positive effect on the mental state of those using it and has the ability to improve their activity. Therefore computational humor deserves particular attention, as it has the potential of changing computers into a creative and motivational tool for human activity [Stock et al., 2002, Nijholt et al., 2003].

While previous work in computational humor has focused mainly on the task of humor generation [Stock and Strapparava, 2003, Binsted and Ritchie, 1997], very few attempts have been made to develop systems for automatic humor recognition [Taylor and Mazlack, 2004]. This is not surprising, since, from a computational perspective, humor recognition appears to be significantly more subtle and difficult than humor generation.

In this paper, we explore the applicability of computational approaches to the recognition of verbally expressed humor. In particular, we investigate whether text classification techniques are a viable approach to distinguish between humorous and non-humorous text, and we bring empirical evidence in support of this hypothesis through experiments performed on very large data sets.

Since a deep comprehension of humor in all of its aspects is probably too ambitious and beyond the existing computational capabilities, we chose to restrict our investigation only to the type of humor found in the *one-liners*. A one-liner is a short sentence with comic effects and an interesting linguistic structure: simple syntax, deliberate use of rhetoric devices (e.g. alliteration and/or rhyme), and frequent use of creative language constructions meant to attract the readers' attention. While longer jokes can have a relatively complex narrative structure, the one-liners must produce the humorous effect "in one shot", with very few words. These characteristics make this type of humor particularly suitable for use in an automatic learning setting, as the humor-producing features are guaranteed to be present in the first (and only) sentence.

We attempt to formulate the humor-recognition problem as a traditional machine learning task, and feed positive (humorous) and negative (non-humorous) examples to an automatic classifier. The humorous data set consists of one-liners collected from the Web using an automatic bootstrapping process. The non-humorous data is selected such that it is structurally and stylistically similar to the one-liners. Specifically, we use three different negative data sets: (1) Reuters news titles; (2) proverbs; and (3) sentences from the British National Corpus (BNC). The classification results achieved with these data sets are very encouraging, with accuracy figures ranging from 77.84% (one-liners/BNC) to 96.89% (one-liners/Reuters). Regardless of the non-humorous data set playing the role of negative examples, the performance of the automatically learned humor-recognizer is always significantly better than apriori known baselines.

The remainder of the paper is organized as follows. We first describe the humorous and non-humorous data sets, and provide details on the Web-based bootstrapping process employed in building a very large collection of one-liners. We then show experimental results obtained on these data sets using two different text classifiers. Finally, we conclude with a discussion and directions for future work.

## Humorous and Non-humorous Data Sets

To test our hypothesis that automatic classification techniques are a viable approach to humor recognition, we needed in the first place a data set consisting of both humorous and non-humorous examples. Once constructed, such data sets can be used to automatically *learn* computational models for humor recognition, and at the same time *evaluate* the performance of such models.

While there is plenty of non-humorous data that can play the role of negative examples, it is significantly harder to build a very large and at the same time sufficiently "clean" data set of humorous examples. We conducted our experiments using two sets of humorous (positive) examples, each of them maximizing a different aspect of the data: (1) *Data quality*: a small set of manually assembled data, guaranteed to be "clean", and (2) *Data quantity*: a very large set of examples automatically collected, which is likely to also include noisy examples.

## Humorous Data

For reasons outlined earlier, we restrict our attention to one-liners, short humorous sentences that have the characteristic of producing a comic effect in very few words (usually 15 or less). The one-liners humor style is illustrated in Table 1, which shows three examples of such one-sentence jokes.

It is well-known in the machine learning community that large amounts of training data have the potential of improving the accuracy of the learning process, and at the same time providing insights into how increasingly larger data sets can affect the classification precision. However, the manual construction of a very large one-liner data set may be problematic, as most Web sites and mailing lists that make available such jokes do not usually list more than 50–100 one-liners. To circumvent this problem, we designed and implemented an automatic bootstrapping approach, which was used to automatically construct a very large collection of 20,000 one-liners.

The main goal of the bootstrapping algorithm is to automatically collect a large number of one-liners, starting with a short *seed* list, consisting of few (ten or less) one-liners manually identified. The bootstrapping process is illustrated in Figure 1. Starting with the seed set, the algorithm automatically identifies a list of webpages that include at least one of the seed one-liners, via a simple search performed with a Web search engine[1]. Next, the webpages found in this way are parsed, and additional one-liners are automatically identified and added to the seed set. The process is then repeated several times, until enough one-liners are collected.

An important aspect of any bootstrapping algorithm is the set of constraints used to steer the process and prevent as much as possible the addition of noisy entries. The one-liner bootstrapping algorithm is guided by two constraints: (1) a *thematic* constraint applied on the content of each webpage; and (2) a *structural* constraint, exploiting HTML annotations indicating "stylistically" similar text.

The first constraint is implemented using a set of keywords of which at least one has to appear in the URL of a retrieved webpage, thus potentially limiting the content of the webpage to a theme related to that keyword. The set of keywords used in the current implementation con-



Figure 1: Web-based bootstrapping of one-liners.

sists of six words that explicitly indicate humor-related content: *oneliner*, *one-liner*, *humor*, *humour*, *joke*, *funny*. For example, `http://www.berro.com/Jokes` or `http://www.mutedfaith.com/funny/life.htm` are the URLs of two webpages that satisfy this first constraint.

The second constraint is designed to exploit the HTML structure of the webpages, in an attempt to identify enumerations of texts that include the seed one-liner. This is based on the hypothesis that enumerations typically include stylistically similar texts, and thus a list including the seed one-liner is very likely to include additional one-line jokes. For instance, if a seed one-liner is found in a webpage preceded by the HTML tag `<li>`[2], other lines found in the same enumeration preceded by the same tag are also likely to be one-liners.

Two iterations of the bootstrapping process, started with a small seed set of ten one-liners, resulted into a large set of about 24,000 one-liners. After removing the duplicates, we were left with a final set of approximately 20,000 one-liners, which were used in the humor-recognition experiments.

## Non-humorous Data

To construct the set of negative examples required by the humor-recognition models, we tried to identify collections of sentences that were non-humorous, but similar in structure and composition to the one-liners. This similarity was sought mainly for the purpose of making

---

[1]Current experiments rely on Google, but other search engines can be used to the same effect. A maximum of 100 candidate URLs are retrieved in return to a search.
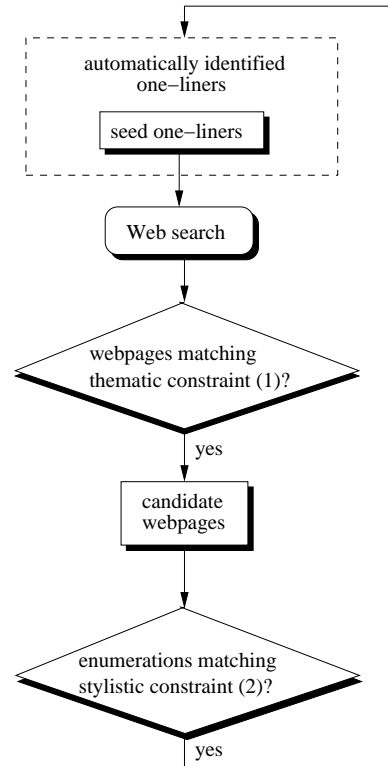
[2]The HTML tag `<li>` stands for "list item."

| One-liners |
|---|
| Take my advice; I don't use it anyway. |
| I get enough exercise just pushing my luck. |
| Beauty is in the eye of the beer holder. |

| Reuters titles |
|---|
| Trocadero expects tripling of revenues. |
| Silver fixes at two-month high, but gold lags. |
| Oil prices slip as refiners shop for bargains. |

| BNC sentences |
|---|
| They were like spirits, and I loved them. |
| I wonder if there is some contradiction here. |
| The train arrives three minutes early. |

| Proverbs |
|---|
| Creativity is more important than knowledge. |
| Beauty is in the eye of the beholder. |
| I believe no tales from an enemy's tongue. |

Table 1: Sample examples of one-liners, Reuters titles, BNC sentences, and proverbs.

the humor-recognition task more difficult and thus more real. We do not want the automatic classifiers to learn to distinguish between humorous and non-humorous examples based simply on text length or vocabulary differences. Instead, we seek to enforce the classifiers to identify humor-specific features, by supplying them with negative examples similar in most of their aspects to the positive examples, but different in their comic effect.

Structural similarity was enforced by requiring that each example in the non-humorous data set follows the same length restriction as the one-liners: one sentence with an average length of 10–15 words. Composition similarity is sought by trying to identify examples similar to the one-liners with respect to their creativity and intent.

We tested three different sets of negative examples:

1. *Reuters* titles, extracted from news articles published in the Reuters newswire over a period of one year (8/20/1996 – 8/19/1997) [Lewis et al., 2004]. The titles consist of short sentences with simple syntax, and are often phrased to catch the readers' attention (an effect similar to the one rendered by one-liners).

2. *Proverbs* manually extracted from an *"English proverb collection."* Proverbs are sayings that transmit, usually in one short sentence, important facts or experiences that are considered true by many people. Their property of being condensed, but memorable sayings make them very similar to the one-liners. In fact, some one-liners attempt to imitate proverbs, but with a comic effect, as in e.g. *"Beauty is in the eye of the beer holder"*, derived from *"Beauty is in the eye of the beholder"*.

3. *British National Corpus (BNC)* sentences, which were

selected at random from the BNC corpus, covering different styles, genres and domains. Unlike the Reuters titles or the proverbs, the BNC sentences have typically no added creativity and no specific intent. However, we decided to add this set of negative examples to our experimental setting, in order to observe the level of difficulty of a humor-recognition task when performed with respect to simple text.

Table 1 shows three examples from each data set, to illustrate their structure and composition.

### The "400HS" and "40000HS" Data Sets

To summarize, two data sets were built and used in the experiments: (1) a small set that emphasizes the *quality* aspect of the data, for which the one-liners were manually selected; and (2) a very large set automatically extracted using a Web-based bootstrapping process, emphasizing the *quantity* aspect of the data, including a small fraction of potentially noisy examples.

- The **"400HS"** data set. In this set, the positive examples consist of 200 one-liners that were manually collected, and thus are guaranteed to be "clean" humorous examples. The set of negative examples consist of one of the following sets: (1) 200 Reuters titles; (2) 200 sentences randomly selected from BNC; (3) 200 proverbs.

- The **"40000HS"** data set. The positive examples in this set consist of 20,000 one-liners automatically identified on the Web using the bootstrapping method illustrated earlier. Since the collection process was automatic, noisy entries are also possible. Manual verification of a randomly selected sample of 200 one-liners resulted into the identification of 18 noisy entries, indicating an average of 9% potential noise in the data set, which is within reasonable limits. The negative examples are drawn from: (1) Reuters titles; or (2) BNC sentences. Since the collection of proverbs that we could obtain was relatively small, this type of negative examples was not included in the large data experiments.

## Algorithms for Text Classification

There is a large body of algorithms previously tested on text classification problems, due also to the fact that text categorization is one of the testbeds of choice for machine learning. In the classification experiments we present here, we compare results obtained with two frequently used text classifiers, Naive Bayes and Support Vector Machines, selected based on their performance in previously reported work, and for the diversity of their learning methodologies.

**Naive Bayes.** The basic idea in a Naive Bayes text classifier is to estimate the probability of a category

given a document using joint probabilities of words and documents. Naive Bayes assumes word independence, which means that the conditional probability of a word given a category is assumed to be independent of the conditional probability of other words given the same category. Despite this simplification, Naive Bayes classifiers perform reasonably well on text classification [Yang and Liu, 1999]. While there are several versions of Naive Bayes classifiers (variations of multinomial and multivariate Bernoulli), we use the multinomial model [McCallum and Nigam, 1998], which was shown to be more effective.

**Support Vector Machines.** Support Vector Machines (SVM) are binary classifiers that attempt to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin [Vapnik, 1995]. Applications of SVM classifiers to text categorization led to some of the best results reported in the literature [Joachims, 1998].

## Experimental Results

The major goal of the studies reported in this paper was to test whether automatic classification techniques can be successfully applied to the task of humor-recognition. To this end, several experiments were conducted to gain insights into various aspects of an automatic humor identification task: classification accuracy, learning rates, impact of the type of negative data used in the learning process, and impact of the classification methodology.

In all the experiments, the evaluation is performed using stratified ten-fold cross validations, to guarantee accurate precision estimates.

Due to the methodology used in building the data sets (equal distribution between positive and negative examples), the baseline for all the experiments is 50%, which represents the classification accuracy obtained if a default label of "humorous" (or "non-humorous") would be assigned by default to all the examples in the data set.

| Classifier | One-liners Reuters | One-liners BNC | One-liners Proverbs |
|---|---|---|---|
| Naive Bayes | 89.75% | 56.75% | 68.50% |
| SVM | 84.75% | 63.75% | 70.00% |

Table 2: Classification accuracy for the **"400HS"** set.

Table 2 shows results obtained on the **"400HS"** data set, for the three different sets of negative examples (Reuters, BNC, Proverbs), using the Naive Bayes and SVM text classifiers. Similar classification results, but this time for the larger **"40000HS"** data set, are shown in Table 3, again with different sets of negative examples (Reuters and BNC), and two different classifiers. Learning curves for this large data set are plotted in Figures 2 and 3.

| Classifier | One-liners Reuters | One-liners BNC |
|---|---|---|
| Naive Bayes | 96.89% | 73.62% |
| SVM | 96.09% | 77.84% |

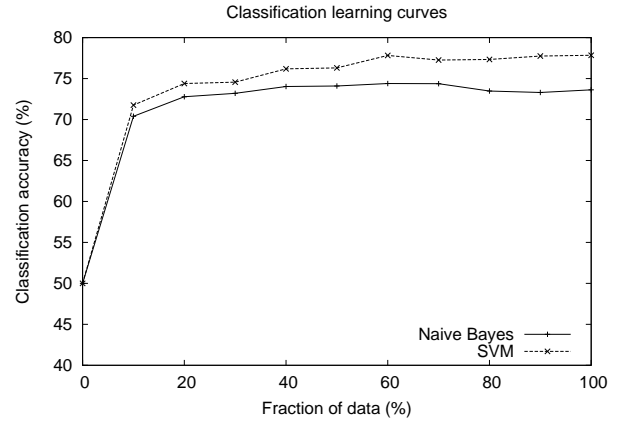Table 3: Classification accuracy for the **"40000HS"** set.

Figure 2: Classification learning curves for the **"40000HS"** (one-liners/BNC) data set.

## Discussion

The results obtained in the automatic classification experiments reveal the fact that computational approaches represent a viable solution for the task of humor-recognition, and good performance can be achieved using standard text classification techniques.

When a clean, manually constructed data set is used (**"400HS"**), a relatively small number of examples (400) was enough to achieve classification accuracies ranging from 56.75% (one-liners/BNC) to 89.75% (one-liners/Reuters), representing a significant improvement over the baseline of 50%.

Although the results obtained in this first set of experiments were already satisfactory, a significantly larger data set was required in order to gain additional insights into the advantages and potential limitations of this automatic classification approach to humor recognition. In addition to accuracy figures, we were also interested in the variation of classification performance with respect to data size, which is an aspect particularly relevant for directing future research. Depending on the shape of the learning curves, one could decide to concentrate future work either on the acquisition of larger data sets, or toward the identification of more sophisticated features. In order to perform these analyses, a very large data set of humorous and non-humorous texts was required, and we used the **"40000HS"** data set automatically bootstrapped from the Web.

For this large, even if noisier data set, the overall performance increased significantly to accuracy figures
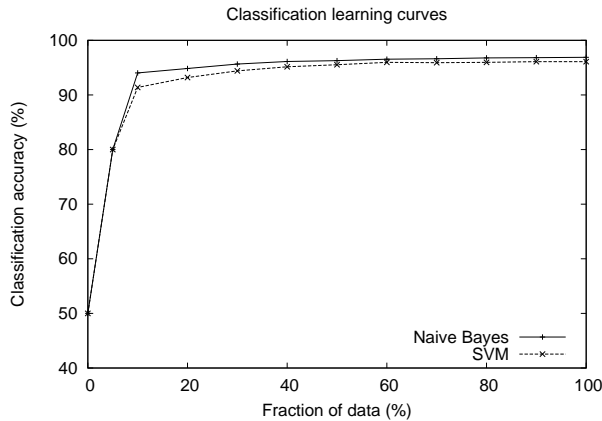
Figure 3: Classification learning curves for the **"40000HS"** (one-liners/Reuters) data set.

ranging from 77.84% (one-liners/BNC) to 96.89% (one-liners/Reuters), representing a major improvement over both the default baseline of 50%, and over the classification results obtained with the **"400HS"** data set.

To evaluate the effect of data *quality* on the classification performance, we also ran an experiment where 400 examples were randomly selected from the large **"40000HS"** corpus, while maintaining the equal distribution between positive and negative examples. This new corpus is therefore of comparable size and characteristics with the **"400HS"** corpus, but of different quality. Table 4 shows the results obtained on this new data set. Comparing the figures in this table with those listed in Table 2, it is clear that data quality can have an important impact on the humor-recognition performance. However, larger, even if noisier, data sets have the ability to outweigh this effect, as shown in the results listed in Table 3.

| Classifier | One-liners Reuters | One-liners BNC |
|---|---|---|
| Naive Bayes | 85.37% | 55.00% |
| SVM | 83.75% | 55.75% |

Table 4: Classification accuracy for a subset of 400 examples from the **"40000HS"** data set.

The learning curves in Figures 2 and 3 show that regardless of the type of negative data and the classifier used, there is significant learning until about 60% of the data (i.e. about 10–12,000 positive examples, and the same number of negative examples). The rather steep ascent of the curve, especially in the first part of the learning, suggests that humorous and non-humorous texts represent well distinguishable types of data.

An interesting effect can be noticed toward the end of the learning, where for both Naive Bayes and SVM

the curve becomes completely flat (One-liners/Reuters), or it even has a slight drop (One-liners/BNC). This is probably due to the presence of noise in the data set, which starts to become visible for very large data sets [3].

The plateau reached at the end of the learning curves is also suggesting that more data is not likely to help improve the quality of an automatic humor-recognizer. Instead, more sophisticated features that go beyond simple bag-of-words analysis are probably required. The type of features to use is a matter of future investigations, and will probably include humor-specific features previously proposed in linguistic studies on humor such as [Bucaria, 2004].

Another interesting result refers to the effect achieved with the various types of negative data. Despite our initial intuition that one-liners are most similar to other creative texts (e.g. Reuters titles, or the sometimes almost identical proverbs), and thus the learning task would be more difficult in relation to these data sets, comparative experimental results reveal the fact that in fact it is more difficult to distinguish humor with respect to regular text (e.g. BNC sentences).

## Related Work

While humor is relatively well studied in scientific fields such as linguistics (e.g. [Attardo, 1994]) and psychology (e.g. [Freud, 1905, Ruch, 2002]), to date there is only a limited number of research contributions made toward the construction of computational humour prototypes.

One of the first attempts is perhaps the work described in [Binsted and Ritchie, 1997], where a formal model of semantic and syntactic regularities was devised, underlying some of the simplest types of puns (*punning riddles*). The model was then exploited in a system called JAPE that was able to automatically generate amusing puns.

Another humor-generation project was the HA-HAcronym project [Stock and Strapparava, 2003], whose goal was to develop a system able to automatically generate humorous versions of existing acronyms, or to produce a new amusing acronym constrained to be a valid vocabulary word, starting with concepts provided by the user. The comic effect was achieved mainly by exploiting incongruity theories (e.g. finding a religious variation for a technical acronym).

Another related work, devoted this time to the problem of humor comprehension, is the study reported in [Taylor and Mazlack, 2004], focused on a very restricted type of wordplays, namely the "Knock-Knock" jokes. The goal of the study was to evaluate to what extent wordplay can be automatically identified in "Knock-Knock" jokes, and if such jokes can be reliably recognized from other non-humorous text. The algorithm was based

---

[3]We also like to think of this behavior as if the computer is losing its sense of humor after an overwhelming number of jokes, in a way similar to humans when they get bored and stop appreciating humor after hearing too many jokes.

on automatically extracted structural patterns and on heuristics heavily based on the peculiar structure of this particular type of jokes. While the wordplay recognition gave satisfactory results, the identification of jokes containing such wordplays turned out to be significantly more difficult.

## Conclusion

The creative genres of natural language have been traditionally considered outside the scope of any computational treatment. In particular humor, because of its puzzling nature, has received little attention from computational linguists. However, given the importance of humor in our everyday life, and the increasing importance of computers in our work and entertainment, we believe that studies related to computational humor will become increasingly important.

In this paper, we showed that automatic classification techniques can be successfully applied to the task of humor-recognition. Experimental results obtained on very large data sets showed that learning approaches can be efficiently used to distinguish between humorous and non-humorous texts, with significant improvements observed over apriori known baselines. To our knowledge, this is the first result of this kind reported in the literature, as we are not aware of any previous work investigating the interaction between humor and machine learning.

Moreover, we have also showed that it is possible to bootstrap a very large and relatively clean corpus that falls under a certain genre (e.g. humor), starting with a handful of manually selected seeds, and using constraints based on document structural information and simple thematic clues. Although current experiments relying on this technique have focused on building a collection of humorous texts, we believe that this Web-based bootstrapping method is not limited to one-liners, but it can be equally well applied to other creative genres.

Finally, through the analysis of learning curves plotting the classification performance with respect to data size, we showed that the accuracy of the automatic humor-recognizer stops improving after a certain number of examples. Given that automatic humor-recognition is a rather understudied problem, we believe that this is an important result, as it gives us insights into potentially productive directions for future work. The flattened shape of the curves toward the end of the learning process suggests that rather than focusing on gathering more data, future work should concentrate on identifying more sophisticated humor-specific features, e.g. semantic oppositions, ambiguity, and others. We plan to address these aspects in future research.

## References

[Attardo, 1994] Attardo, S. (1994). *Linguistic Theory of Humor*. Mouton de Gruyter, Berlin.

[Binsted and Ritchie, 1997] Binsted, K. and Ritchie, G. (1997). Computational rules for punning riddles. *Humor*, 10(1).

[Bucaria, 2004] Bucaria, C. (2004). Lexical and syntactic ambiguity as a source of humor. *Humor*, 17(3).

[Freud, 1905] Freud, S. (1905). *Der Witz und Seine Beziehung zum Unbewussten*. Deutike, Vienna.

[Joachims, 1998] Joachims, T. (1998). Text categorization with Support Vector Machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*.

[Lewis et al., 2004] Lewis, D., Yang, Y., Rose, T., and Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397.

[McCallum and Nigam, 1998] McCallum, A. and Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*.

[Nijholt et al., 2003] Nijholt, A., Stock, O., Dix, A., and Morkes, J., editors (2003). *Proceedings of CHI-2003 workshop: Humor Modeling in the Interface*, Fort Lauderdale, Florida.

[Ruch, 2002] Ruch, W. (2002). Computers with a personality? lessons to be learned from studies of the psychology of humor. In *[Stock et al., 2002]*.

[Stock and Strapparava, 2003] Stock, O. and Strapparava, C. (2003). Getting serious about the development of computational humour. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico.

[Stock et al., 2002] Stock, O., Strapparava, C., and Nijholt, A., editors (2002). *Proceedings of the The April Fools Day Workshop on Computational Humour (TWLT20)*, Trento.

[Taylor and Mazlack, 2004] Taylor, J. and Mazlack, L. (2004). Computationally recognizing wordplay in jokes. In *Proceeding of CogSci 2004*, Chicago.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

[Yang and Liu, 1999] Yang, Y. and Liu, X. (1999). A reexamination of text categorization methods. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*.