Predicting the Humorousness of Tweets Using Gaussian Process Preference Learning

Identificando el humor de tuits utilizando el aprendizaje de preferencias basado en procesos gaussianos

Tristan Miller[1], Erik-Lân Do Dinh[2], Edwin Simpson[2], Iryna Gurevych[2]
[1]Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6, 1010 Vienna, Austria
tristan.miller@ofai.at
[2]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt, Germany
https://www.ukp.tu-darmstadt.de/

Abstract: Most humour processing systems to date make at best discrete, coarse-grained distinctions between the comical and the conventional, yet such notions are better conceptualized as a broad spectrum. In this paper, we present a probabilistic approach, a variant of Gaussian process preference learning (GPPL), that learns to rank and rate the humorousness of short texts by exploiting human preference judgments and automatically sourced linguistic annotations. We apply our system, which is similar to one that had previously shown good performance on English-language one-liners annotated with pairwise humorousness annotations, to the Spanish-language data set of the HAHA@IberLEF2019 evaluation campaign. We report system performance for the campaign's two subtasks, humour detection and funniness score prediction, and discuss some issues arising from the conversion between the numeric scores used in the HAHA@IberLEF2019 data and the pairwise judgment annotations required for our method.
Keywords: Computational humour, humour, Gaussian process preference learning, GPPL, best–worst scaling

Resumen: Actualmente la mayoría de los sistemas de procesamiento de humor hacen, en el mejor de los casos, distinciones discretas y granualeres entre lo cómico y lo convencional. Sin embargo, dichos conceptos se conciben mejor en un espectro más amplio. Este artículo presenta un método probabilístico, un modo de preferencias de aprendizaje basadas en un proceso gaussiano (GPPL), que aprende a clasificar y calificar el humor de textos cortos explotando juicios de preferencia humana y anotaciones lingüísticas generadas en forma automática. Nuestro sistema es similar a uno que previamente había demostrado un buen desempeño en frases en inglés anotadas con anotaciones humorísticas por pares y lo aplicamos a la colección de datos en español de la campaña de evaluación HAHA@IberLEF2019. En este trabajo reportamos el desempeño del sistema para dos subtareas de la campaña: detección de humor y predicción de puntaje de diversión. También presentamos algunos problemas que surgen de la conversión entre los puntajes numéricos utilizados en los datos HAHA@IberLEF2019 y las anotaciones de juicio de pares de documentos requeridas para nuestro método.
Palabras clave: Humor computacional, humor, preferencias de aprendizaje, proceso gaussiano, GPPL, mejor-peor escala

Tristan Miller, Erik-Lân Do Dinh, Edwin Simpson, Iryna Gurevych

1 Introduction

Humour is an essential part of everyday communication, particularly in social media (Holton and Lewis, 2011; Shifman, 2013), yet it remains a challenge for computational methods. Unlike conventional language, humour requires complex linguistic and background knowledge to understand, which are difficult to integrate with NLP methods (Hempelmann, 2008).

An important step in the automatic processing of humour is to recognize its presence in a piece of text. However, its intensity may be present or perceived to varying degrees to its human audience (Bell, 2017). This level of appreciation (i.e., *humorousness* or equivalently *funniness*) can vary according to the text's content and structural features, such as nonsense or disparagement (Carretero-Dios, Pérez, and Buela-Casal, 2010) or, in the case of puns, contextual coherence (Lippman and Dunn, 2000) and the cognitive effort required to recover the target word (Hempelmann, 2003, pp. 123–124).

While previous work has considered mainly binary classification approaches to humorousness, the HAHA@IberLEF2019 shared task (Chiruzzo et al., 2019) also focuses on its gradation. This latter task is important for downstream applications such as conversational agents or machine translation, which must choose the correct tone in response to humour, or find appropriate jokes and wordplay in a target language. The degree of creativeness may also inform an application whether the semantics of a joke can be inferred from similar examples.

This paper describes a system designed to carry out both subtasks of the HAHA@IberLEF2019 evaluation campaign: binary classification of tweets as humorous or not humorous, and the quantification of humour in those tweets. Our system employs a Bayesian approach—namely, a variant of Gaussian process preference learning (GPPL) that infers humorousness scores or rankings on the basis of manually annotated pairwise preference judgments and automatically annotated linguistic features. In the following sections, we describe and discuss the background and methodology of our system, our means of adapting the HAHA@IberLEF2019 data to work with our system, and the results of our system evaluation on this data.

2 Background

Pairwise comparisons can be used to infer rankings or ratings by assuming a *random utility model* (Thurstone, 1927), meaning that the annotator chooses an instance (from a pair or list of instances) with probability p, where p is a function of the *utility* of the instance. Therefore, when instances in a pair have similar utilities, the annotator selects one with a probability close to 0.5, while for instances with very different utilities, the instance with higher utility will be chosen consistently. The random utility model forms the core of two popular preference learning models, the Bradley–Terry model (Bradley and Terry, 1952; Luce, 1959; Plackett, 1975), and the Thurstone–Mosteller model (Thurstone, 1927; Mosteller, 1951). Given this model and a set of pairwise annotations, probabilistic inference can be used to retrieve the latent utilities of the instances.

Besides pairwise comparisons, a random utility model is also employed by MaxDiff (Marley and Louviere, 2005), a model for best–worst scaling (BWS), in which the annotator chooses the best and worst instances from a set. While the term "best–worst scaling" originally applied to the data collection technique (Finn and Louviere, 1992), it now also refers to models such as MaxDiff that describe how annotators make discrete choices. Empirical work on BWS has shown that Max-Diff scores (instance utilities) can be inferred using either maximum likelihood or a simple counting procedure that produces linearly scaled approximations of the maximum likelihood scores (Flynn and Marley, 2014). The counting procedure defines the score for an instance as the fraction of times the instance was chosen as best, minus the fraction of times the instance was chosen as worst, out of all comparisons including that instance (Kiritchenko and Mohammad, 2016). From this point on, we refer to the counting procedure as BWS, and apply it to the tasks of inferring scores from pairwise annotations for funniness.

Gaussian process preference learning (GPPL) (Chu and Ghahramani, 2005), a Thurstone–Mosteller–based model that accounts for the features of the instances when inferring their scores, can make predictions for unlabelled instances and copes better with sparse pairwise labels. GPPL uses Bayesian inference, which has been shown to cope better with sparse and noisy data (Xiong, Barash,

and Frey, 2011; Titov and Klementiev, 2012; Beck, Cohn, and Specia, 2014; Lampos et al., 2014), including disagreements between multiple annotators (Cohn and Specia, 2013; Simpson et al., 2015; Felt, Ringger, and Seppi, 2016; Kido and Okamoto, 2017). Through the random utility model, GPPL handles disagreements between annotators as noise, since no instance in a pair has a probability of one of being selected.

Given a set of pairwise labels, and the features of labelled instances, GPPL can estimate the posterior distribution over the utilities of any instances given their features. Relationships between instances are modelled by a Gaussian process, which computes the covariance between instance utilities as a function of their features (Rasmussen and Williams, 2006). Since typical methods for posterior inference (Nickisch and Rasmussen, 2008) are not scalable (the computational complexity is $\mathcal{O}(n^3)$, where n is the number of instances), we use a scalable method for GPPL that permits arbitrarily large numbers of instances and pairs (Simpson and Gurevych, 2018). This method uses stochastic variational inference (Hoffman et al., 2013), which limits computational complexity by substituting the instances for a fixed number of *inducing points* during inference.

The GPPL method has already been applied with good results to ranking arguments by convincingness (which, like funniness, is an abstract linguistic property that is hard to quantify directly) and to ranking English-language one-liners by humorousness (Simpson and Gurevych, 2018; Simpson et al., 2019). In these two tasks, GPPL was found to outperform SVM and BiLSTM regression models that were trained directly on gold-standard scores, and to outperform BWS when given sparse training data, respectively. We therefore elect to use GPPL on the Spanish-language Twitter data of the HAHA@IberLEF2019 shared task.

In the interests of replicability, we freely release the code for running our GPPL system, including the code for the data conversion and subsampling process detailed in §3.2.[1]

---

[1] https://github.com/UKPLab/haha2019-GPPL

3 Experiments

3.1 Tasks

The HAHA@IberLEF2019 evaluation campaign consists of two tasks. Task 1 is humour detection, where the goal is to predict whether or not a given tweet is humorous, as determined by a gold standard of binary, human-sourced annotations. Systems are scored on the basis of accuracy, precision, recall, and F-measure. Task 2 is humorousness prediction, where the aim is to assign each funny tweet a score approximating the average funniness rating, on a five-point scale, assigned by a set of human annotators. Here system performance is measured by root-mean-squared error (RMSE). For both tasks, the campaign organizers provide a collection of 24 000 manually annotated training examples. The test data consists of a further 6000 tweets whose gold-standard annotations were withheld from the participants.

3.2 Data Preparation

For each of the 24 000 tweets in the HAHA@IberLEF2019 training data, the task organizers asked human annotators to indicate whether the tweet was humorous, and if so, how funny they found it on a scale from 1 ("not funny") to 5 ("excellent"). This is essentially the same annotation scheme used for the first version of the corpus (Castro et al., 2018) which was used in the previous iteration of HAHA (Castro, Chiruzzo, and Rosá, 2018). As originally distributed, the training data gives the text of each tweet along with the number of annotators who rated it as "not humour", "1", "2", "3", "4", and "5". For the purposes of Task 1, tweets in the positive class received at least three numerical annotations and at least five annotations in total; tweets in the negative class received at least three "not humour" annotations, though possibly fewer than five annotations in total. Only those tweets in the positive class are used in Task 2.

This ordinal data cannot be used as-is with our GPPL system, which requires as input a set of preference judgments between pairs of instances. To work around this, we converted the data into a set of ordered pairs of tweets such that the first tweet has a lower average funniness score than the second. (We consider instances in the negative class to have an average funniness score of 0.) While an exhaustive set of pairings would contain 575 976 000 pairs

(minus the pairs in which both tweets have the same score), we produced only $10\,730\,229$ pairs, which was the minimal set necessary to accurately order the tweets. For example, if the original data set contained three tweets A, B, and C with average funniness scores 5.0, 3.0, and 1.0, respectively, then our data would contain the pairs (C, B) and (B, A) but not (C, A). To save memory and computation time in the training phase, we produced a random subsample such that the number of pairs where a given tweet appeared as the funnier one was capped at 500. This resulted in a total of $485\,712$ pairs. In a second configuration, we subsampled up to 2500 pairs per tweet. We used a random 60% of this set to meet memory limitations, resulting in $686\,098$ pairs.

With regards to the tweets' textual data, we do only basic tokenization as preprocessing. For lookup purposes (synset lookup; see §3.3), we also lemmatize the tweets.

3.3 Experimental Setup

As we adapt an existing system that works on English data (Simpson et al., 2019), we generally reuse the features employed there, but use Spanish resources instead. Each tweet is represented by the vector resulting from a concatenation of the following:

- The average of the word embedding vectors of the tweet's tokens, for which we use 200-dimensional pretrained Spanish Twitter embeddings (Deriu et al., 2017).

- The average frequency of the tweet's tokens, as determined by a Wikipedia dump.[2]

- The average word polysemy—i.e., the number of synsets per lemma of the tweet's tokens, as given by the Multilingual Central Repository (MCR 3.0, release 2016) (Gonzalez-Agirre, Laparra, and Rigau, 2012).

Using the test data from the HAHA@IberLEF2018 task (Castro, Chiruzzo, and Rosá, 2018) as a development set, we further identified the following features from the UO_UPV system (Ortega-Bueno et al., 2018) as helpful:

- The heuristically estimated turn count (i.e., the number of tokens beginning with - or --) and binary dialogue heuristic (i.e., whether the turn count is greater than 2).

- The number of hashtags (i.e., tokens beginning with #).

- The number of URLs (i.e., tokens beginning with www or http).

- The number of emoticons.[3]

- The character and token count, as well as mean token length.

- The counts of exclamation marks and other punctuation (.,;?).

We adapt the existing GPPL implementation[4] using the authors' recommended hyperparameter defaults (Simpson and Gurevych, 2018): batch size $|P_i| = 200$, scale hyperparameters $\alpha_0 = 2$ and $\beta_0 = 200$, and the number of inducing points (i.e., the smaller number of data points that act as substitutes for the tweets in the dataset) $M = 500$. The maximum number of iterations was set to 2000. Using these feature vectors, hyperparameter settings, and data pairs, we require a training time of roughly two hours running on a 24-core cluster with 2 GHz CPU cores.

After training the model, an additional step is necessary to transform the GPPL output values to the original funniness range (0, 1–5). For this purpose, we train a Gaussian process regressor which we supply with the output values of the GPPL system as features and the corresponding HAHA@IberLEF2018 test data values as targets. However, this model can still yield results outside the desired range when applied to the GPPL output of the HAHA@IberLEF2019 test data. Thus, we afterwards map too-large and too-small values onto the range boundaries. We further set an empirically determined threshold for binary funniness estimation.

3.4 Results and Discussion

Table 1 reports results for the binary classification setup (Task 1) and the regression task (Task 2). Included in each table are the scores

---

[2] `https://dumps.wikimedia.org/eswiki/ 20190420/eswiki-20190420-pages-articles. xml.bz2`; last accessed on 2019-06-15.

[3] `https://en.wikipedia.org/wiki/List_of_ emoticons\#Western`, Western list; last accessed on 2019-06-15.

[4] `https://github.com/UKPLab/ tacl2018-preference-convincing`

System	Task 1				Task 2
	F_1	Precision	Recall	Accuracy	RMSE
Ismailov (2019)	0.821	0.791	0.852	0.855	0.736
our system	0.660	0.588	0.753	0.698	1.810
baseline	0.440	0.394	0.497	0.505	2.455

Table 1: Results for Task 1 (humour detection) and Task 2 (funniness score prediction)

of our own system, as well as those of the top-performing system (Ismailov, 2019) and a naïve baseline. For Task 1, the naïve baseline makes a random classification for each tweet (with uniform distribution over the two classes); for Task 2, it assigns a funniness score of 3.0 to each tweet.

In the binary classification setup, our system achieved an F-measure of 0.660 on the test data, representing a precision of 0.588 and a recall of 0.753. In the regression task, we achieved RMSE of 1.810. The results are based on the second data subsample (up to 2500 pairs), with the results for the first (up to 500 pairs) being slightly lower. Our results for both tasks, while handily beating those of the naïve baseline, are significantly worse than those reported by some other systems in the evaluation campaign, including of course the winner. This is somewhat surprising given GPPL's very good performance in previous English-language experiments (Simpson et al., 2019).

Unfortunately, our lack of fluency in Spanish and lack of access to the gold-standard scores for the test set tweets precludes us from performing a detailed qualitative error analysis. However, we speculate that our system's less than stellar performance can partly be attributed to the information loss in converting between the numeric scores used in the HAHA@IberLEF2019 tasks and the preference judgments used by our GPPL system. In support of this explanation, we note that the output of our GPPL system is rather uniform; the scores occur in a narrow range with very few outliers. (Figure 1 shows this outcome for the HAHA@IberLEF2018 test data.) Possibly this effect would have been less pronounced had we used a much larger subsample, or even the entirety, of the possible training pairs, though as discussed in §3.2, technical and temporal limitations prevented us from doing so. We also speculate that the Gaussian process regressor we used may not have been the best way of mapping our GPPL scores
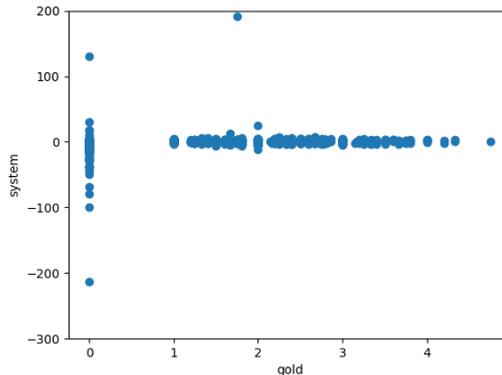

Figure 1: Gold values of the HAHA@IberLEF2018 test data (x axis) vs. the scores assigned by our GPPL system (y axis), before mapping to the expected funniness range using a Gaussian process regressor. The lowest GPPL value (-1400) was removed from the plot to obtain a better visualization.

back onto the task's funniness scale (albeit still better than a linear mapping).

Apart from the difficulties posed by the differences in the annotation and scoring, our system may have been affected by the mismatch between the language resources for most of its features and the language of the test data. That is, while we relied on language resources like Wikipedia and MCR that reflect standardized registers and prestige dialects, the HAHA@IberLEF2019 data is drawn from unedited social media, whose language is less formal, treats a different range of topics, and may reflect a wider range of dialects and writing styles. Twitter data in particular is known to present problems for vanilla NLP systems, at least without extensive cleaning and normalization (W-NUT, 2015). This is reflected in our choice of word embeddings: while we achieved a Spearman rank correlation of $\rho = 0.52$ with the HAHA@IberLEF2018 test data using embeddings based on Twitter data (Deriu et al., 2017), the same system using more

"standard" Wikipedia-/news-/Web-based embeddings[5] resulted in a correlation near zero.

4 Conclusion

This paper has presented a system for predicting both binary and graded humorousness. It employs Gaussian process preference learning, a Bayesian system that learns to rank and rate instances by exploiting pairwise preference judgments. By providing additional feature data (in our case, shallow linguistic features), the method can learn to predict scores for previously unseen items.

Though our system is based on one that had previously achieved good results with rudimentary, task-agnostic linguistic features on two English-language tasks (including one involving the gradation of humorousness), its performance on the Spanish-language Twitter data of HAHA@IberLEF2019 was less impressive. We tentatively attribute this to the information loss involved in the (admittedly artificial) conversion between the numeric annotations used in the task and the preference judgments required as input to our method, and to the fact that we do not normalize the Twitter data to match our linguistic resources.

A possible avenue of future work, therefore, might be to mitigate the data conversion problem. However, as it has been rather convincingly argued, both generally (Thurstone, 1927) and in the specific case of humour assessment (Shahaf, Horvitz, and Mankoff, 2015), that aggregate ordinal rating data should not be treated as interval data, the proper solution here would be to recast the entire task from one of binary classification or regression to one of comparison or ranking. Perhaps the best way of doing this would be to source new gold-standard preference judgments on the data, though this would be an expensive and time-consuming endeavour.[6]

Regardless of the task setup, there are a few further ways our system might be improved. First, we might try normalizing the language of the tweets, and secondly, we might try using additional, humour-specific features, including some of those used in past work as well as those inspired by the prevailing linguistic theories of humour (Attardo, 1994). The benefits of including word frequency also point to possible further improvements using n-grams, TF–IDF, or other task-agnostic linguistic features.

References

Attardo, S. 1994. *Linguistic Theories of Humor*. Mouton de Gruyter, Berlin.

Beck, D., T. Cohn, and L. Specia. 2014. Joint emotion analysis via multi-task Gaussian processes. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1798–1803.

Bell, N. D. 2017. Failed humor. In S. Attardo, editor, *The Routledge Handbook of Language and Humor*. Routledge, New York, pages 356–370.

Bradley, R. A. and M. E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, December.

Carretero-Dios, H., C. Pérez, and G. Buela-Casal. 2010. Assessing the appreciation of the content and structure of humor: Construction of a new scale. *Humor*, 23(3):307–325, August.

Castro, S., L. Chiruzzo, and A. Rosá. 2018. Overview of the HAHA task: Humor analysis based on human annotation at IberEval 2018. In *Proc. Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, volume 2150 of *CEUR Workshop Proceedings*, pages 187–194.

---

[5] https://zenodo.org/record/1410403

[6] We note with appreciation that the upcoming SemEval-2020 task on humour assessment (Hossain et al., 2020) does include a subtask for predicting preference judgments, though it seems the underlying gold-standard data still uses aggregate ordinal data.

Castro, S., L. Chiruzzo, A. Rosá, D. Garat, and G. Moncecchi. 2018. A crowd-annotated Spanish corpus for humor analysis. In *Proc. Sixth International Workshop on Natural Language Processing for Social Media*, pages 7–11.

Chiruzzo, L., S. Castro, M. Etcheverry, D. Garat, J. J. Prada, and A. Rosá. 2019. Overview of HAHA at IberLEF 2019: Humor analysis based on human annotation. In *Proc. Iberian Languages Evaluation Forum*, volume 2421 of *CEUR Workshop Proceedings*, pages 132–144.

Chu, W. and Z. Ghahramani. 2005. Preference learning with Gaussian processes. In *Proc. 22nd International Conference on Machine Learning*, pages 137–144.

Cohn, T. and L. Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 32–42.

Deriu, J., A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hoffmann, and M. Jaggi. 2017. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proc. 26th International World Wide Web Conference*, pages 1045–1052.

Felt, P., E. K. Ringger, and K. D. Seppi. 2016. Semantic annotation aggregation with conditional crowdsourcing models and word embeddings. In *Proc. 26th International Conference on Computational Linguistics*, pages 1787–1796.

Finn, A. and J. J. Louviere. 1992. Determining the appropriate response to evidence of public concern: The case of food safety. *J. Public Policy & Market.*, 11(2):12–25, September.

Flynn, T. N. and A. A. J. Marley. 2014. Best–worst scaling: Theory and methods. In S. Hess and A. Daly, editors, *Handbook of Choice Modelling*. Edward Elgar Publishing, Cheltenham, UK, pages 178–201.

Gonzalez-Agirre, A., E. Laparra, and G. Rigau. 2012. Multilingual Central Repository version 3.0. In *Proc. 8th International Conference on Language Resources and Evaluation*, pages 2525–2529.

Hempelmann, C. F. 2003. *Paronomasic Puns: Target Recoverability Towards Automatic Generation*. Ph.D. thesis, Purdue University, West Lafayette, IN, USA.

Hempelmann, C. F. 2008. Computational humor: Beyond the pun? In V. Raskin, editor, *The Primer of Humor Research*. Mouton de Gruyter, Berlin, pages 333–360.

Hoffman, M. D., D. M. Blei, C. Wang, and J. W. Paisley. 2013. Stochastic variational inference. *J. Mach. Learn. Res.*, 14:1303–1347, May.

Holton, A. E. and S. C. Lewis. 2011. Journalists, social media, and the use of humor on Twitter. *Electron. J. Commun.*, 21(1&2).

Hossain, N., J. Krumm, M. Gamon, and H. Kautz. 2020. SemEval-2020 Task 7: Assessing the funniness of edited news headlines. In *Proc. 14th International Workshop on Semantic Evaluation*. To appear.

Ismailov, A. 2019. Humor analysis based on human annotation challenge at IberLEF 2019: First-place solution. In *Proc. Iberian Languages Evaluation Forum*, volume 2421 of *CEUR Workshop Proceedings*, pages 160–164.

Kido, H. and K. Okamoto. 2017. A Bayesian approach to argument-based reasoning for attack estimation. In *Proc. 26th International Joint Conference on Artificial Intelligence*, pages 249–255.

Kiritchenko, S. and S. M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 811–817.

Lampos, V., N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. 2014. Predicting and characterising user impact on Twitter. In *Proc. 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413.

Lippman, L. G. and M. L. Dunn. 2000. Contextual connections within puns: Effects on perceived humor and memory. *J. Gen. Psychol.*, 127(2):185–197, April.

Luce, R. D. 1959. On the possible psychophysical laws. *Psychol. Rev.*, 66(2):81–95.

Marley, A. A. J. and J. J. Louviere. 2005. Some probabilistic models of best, worst, and best–worst choices. *J. Math. Psychol.*, 49(6):464–480.

Mosteller, F. 1951. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, March.

Nickisch, H. and C. E. Rasmussen. 2008. Approximations for binary Gaussian process classification. *J. of Mach. Learn. Res.*, 9:2035–2078, October.

Ortega-Bueno, R., C. E. Muñiz Cuza, J. E. Medina Pagola, and P. Rosso. 2018. UO_UPV: Deep linguistic humor detection in Spanish social media. In *Proc. Third Workshop on Evaluation of Human Language Technologies for Iberian Languages*, volume 2150 of *CEUR Workshop Proceedings*, pages 203–213.

Plackett, R. L. 1975. The analysis of permutations. *J. Royal Stat. Soc. Ser. C (Appl. Stat.)*, 24(2):193–202.

Rasmussen, C. E. and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA.

Shahaf, D., E. Horvitz, and R. Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1065–1074.

Shifman, L. 2013. *Memes in Digital Culture.* MIT Press, Cambridge, MA, USA.

Simpson, E., E.-L. Do Dinh, T. Miller, and I. Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728.

Simpson, E. and I. Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Trans. Assoc. Comput. Ling.*, 6:357–371.

Simpson, E. D., M. Venanzi, S. Reece, P. Kohli, J. Guiver, S. J. Roberts, and N. R. Jennings. 2015. Language understanding in the wild: Combining crowdsourcing and machine learning. In *Proc. 24th International Conference on World Wide Web*, pages 992–1002.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychol. Rev.*, 34(4):273–286.

Titov, I. and A. Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *Proc. 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–22.

2015. *Proc. ACL 2015 Workshop on Noisy User-generated Text.* Association for Computational Linguistics.

Xiong, H. Y., Y. Barash, and B. J. Frey. 2011. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinform.*, 27(18):2554–2562, September.