Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5956–5962
Marseille, 20-25 June 2022

Please, Don't Forget the Difference and the Confidence Interval when Seeking for the State-of-the-Art Status

Yves Bestgen
Laboratoire d'analyse statistique des textes
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

This paper argues for the widest possible use of bootstrap confidence intervals for comparing NLP system performances instead of the state-of-the-art status (SOTA) and statistical significance testing. Their main benefits are to draw attention to the difference in performance between two systems and to help assessing the degree of superiority of one system over another. Two cases studies, one comparing several systems and the other based on a K-fold cross-validation procedure, illustrate these benefits.

Keywords: NLP system evaluation, statistical significance testing, bootstrap confidence interval, Fisher-Pitman test, SOTA, resampling approach

1. Introduction

With the explosion of NLP studies, more and more systems are developed and evaluated using more and more shared tasks and freely available datasets. Comparing systems has therefore become an essential question (Berg-Kirkpatrick et al., 2012; Dodge et al., 2019). Very often, it boils down to ranking systems by performance like in leaderboards or determining whether the new system scores better than the previous better one, the widely discussed "glorification of benchmarks" and state-of-the-art status (SOTA) (Dehghani et al., 2021). The major advantage of this ranking criterion is that it only requires the performance level of each system.

Using simple ranking has been criticized because the superiority of one system over another may result from chance alone rather than from any real difference (Koehn, 2004). The answer is obviously statistical significance testing, whose function is to reject the null hypothesis according to which there is no difference in the population to which the evaluation set belong. Unlike SOTA, it requires access to the predictions of the models to be compared. Yet, it has been widely emphasized, but more often in other fields than NLP (Cohen, 1994; Ioannidis, 2019; Kim and Robinson, 2019; Rao and Lovric, 2016; Schmidt and Hunter, 1997), that null-hypothesis significance testing presents many problems. When comparing NLP systems, one of the most important issues is that a statistical test leads to dichotomizing the results of a study in significant vs. not significant, while statistically significant does not mean important. For most statistical tests, any difference, however small it may be and therefore as unimportant as it is, can be declared significant with enough data. This is for example the case when a chi-square test is (wrongly) used to compare the frequency of words or lexical bundles in large corpora: the smallest difference in frequency gives rise to a statistically highly significant result (Cohen, 1994; Bestgen, 2014; Bestgen, 2020a).

For these reasons, statisticians and researchers in many disciplines have strongly recommended reporting an effect size measure to help evaluating the importance of the result (Claridge-Chang and Assam, 2016; Cohen, 1994; Masson and Loftus, 2003; Saville, 1990; Wasserstein et al., 2019). When comparing two systems, the most obvious measure of importance is the true difference in performance between them, i.e. the difference in performance that would be observed if the systems were compared on all data they could process. Unfortunately, but obviously, this value is unknown. The observed difference is the best estimate available, but it can be more or less far from the true difference. This is where the confidence interval (CI) comes into play.

Conceptually close to the hypothesis test[1], it is, on the basis of the available data, the best guess, at some level of confidence (e.g. 95%), of an interval that contains the true difference between the two compared systems (Ismay and Kim, 2019). More formally, the interpretation of a CI at 95% is as follows: if a large number of samples are extracted from the population in question and a CI is calculated for each of these samples, 95% of them should include the population parameter (Dix, 2020).

Figure 1 shows a graphical representation of a confidence interval. The length of the interval informs about the uncertainty associated with the estimate of the true difference and the size of the gap between the confidence limits and 0 helps to assess the degree of superiority of one system over another. This paper argues for

[1] A CI at 95% provides a test of significance at the .05 threshold since the true difference can be declared different from 0 if the CI does not include this value. However, it leads to dichotomizing the results as significant or not, the opposite of what a CI is done for (Wasserstein et al., 2019).

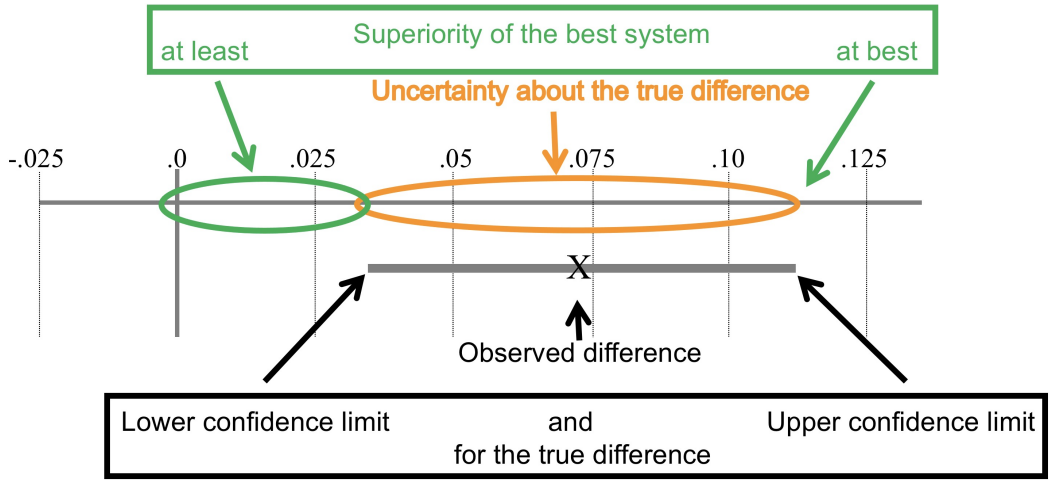

Figure 1: Confidence interval for a difference in accuracy between two systems

the widest possible use of CIs in NLP system comparison.

Calculating CIs for comparing systems is not new in NLP. For example, Koehn (2004), Zhang and Vogel (2004) and Graham et al. (2014) proposed to use bootstrap-style CIs to compare machine translation systems while Bisani and Ney (2004) (see also Liu (2019)) made the same proposal to compare the word error rates of automatic speech recognition systems. However, these studies have emphasized the use of CIs to determine whether a difference is statistically significant and not to deepen the interpretation of the differences in performance between systems. Soboroff (2014) recommended the use of CIs in IR and compared several computation techniques to obtain them. He focused on finding confidence limits for the performance of a system but suggested that they could be useful for comparing several systems. More recently, Lucic et al. (2018) used bootstrap CIs to get a better view on a system performance than that provided by its best result.

To my knowledge, the closest study is that of Berrar and Lozano (2013) who compared the advantages and disadvantages of using significance tests and CIs for the comparison of classifiers. If the arguments presented here are very close to theirs, the proposed approach is much more general by making use of bootstrap CIs.

2. Using Bootstrap CI for Comparing Systems

Since many different performance measures are used in NLP and since the statistical properties of these measures are frequently poorly known, Berg-Kirkpatrick et al. (2012) and Koehn (2004) have recommended to use a resampling approach such as paired bootstrap because it can be applied to any performance metric. Bootstrap CI calculation works as follow (DiCiccio and Efron, 1996).

Given a evaluation set of N instances for which the predictions of the two systems have been obtained, by means of some cross-validation procedure,

1. Compute the observed difference in performance of the two systems on the whole evaluation set, the best available estimate of the true difference.
2. Repeat the following three steps a large number of times[2] to get the bootstrap distribution of the difference.
   - 2*a*) Randomly sample with replacement N instances from the evaluation set to get a bootstrap sample.
   - 2*b*) Calculate the performances of each system on this bootstrap sample.
   - 2*c*) Calculate the difference between these two performances.
3. Use the bootstrap distribution to obtain the CI for the difference.

Different ways of calculating this CI have been proposed such as the percentile method and the bias corrected method with acceleration (BCa), which is considered to be quite accurate, even for skewed distribution, unless the sample size of the evaluation set is very small (Hayden, 2019; Hesterberg et al., 2003).

It is noteworthy that, like a significance test, CIs are affected by the sample size of the evaluation set. *Everything else being equal*, larger sample sizes produce narrower CIs. However, the change in interpretation is very different compared to that of a significance test. A very large sample will produce a very small p-value and the researcher can claim to be very confident about the existence of a real difference *whatever its size*. The larger the sample, the smaller the CI and therefore the more confident the researcher can be about the true difference estimate, but this estimate will be identical. If the observed difference is unimportant, it will remain

[2]For a 95% CI, 1000 to 5000 bootstrap samples are considered as enough (Hesterberg et al., 2003).

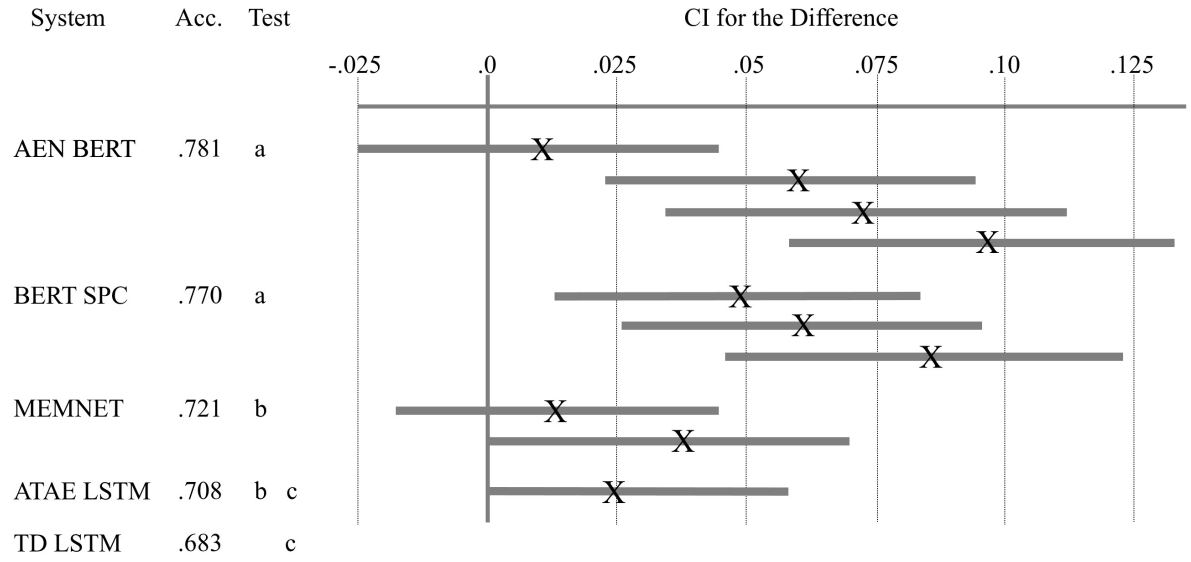

Figure 2: Results of the different ways of comparing systems for ABSA. The systems are ordered vertically according to their accuracy, which is given to the right of the name. The letters (a, b and c) indicate the conclusions of the resampling test ($\alpha = .05$), the performances of two systems having the same letter are not significantly different. The CIs at 95% are given for the difference between each system and those ranked lower, presented in rank order. Thus, the upper CI is for the difference between AEN BERT and BERT SPC and the next one is for for the difference between AEN BERT and MEMNET. Only the first CI of a system is aligned with its name. The *X*s indicate the observed differences.

unimportant (Kim and Robinson, 2019).

A python module for obtaining these CIs is freely available on PyPi and can be installed with *pip install BootCI* (Yves Bestgen, 2022). Details about to use it is available at (`https://github.com/ybestgen/BootCIExpli/`. It is based on the *bootstrap_interval* module by Alexander Neshitov (`https://pypi.org/project/bootstrap-interval/`, MIT Licence, (Alexander Neshitov, 2019)) which implements the bootstrap CI for one sample described in DiCiccio and Efron (1996). It was adapted to the comparison of paired samples. The performance measure can be any measure provided by a standard python library or defined by the user. To facilitate the comparison of the CI approach with the significance test one, a second function implementing the Fisher-Pitman test for paired samples (Berry et al., 2002; Bestgen, 2017; Neuhauser and Manly, 2004) is also provided in the *BootCI* module. As the bootstrap CI, this test is based on a resampling procedure.

3. Two Case Studies

The rest of this paper presents two case studies in the field of emotion mining which illustrate the use of CIs for comparing NLP systems. Explanations, code and data for reproducing these cases studies are freely available. They can be downloaded from `https://github.com/ybestgen/BootCIRealData`. The reported CIs were obtained with the BCa technique, as recommended in the previous section, using 10,000 samples. The same number of samples were used for the Fisher-Pitman tests. In all the analyses, α, the probability of wrongly rejecting the null hypothesis, was set to the usual .05 and the confidence to $(1 - \alpha)$. As CIs are interpreted as providing uncertainty bounds on an estimate and not as significance tests, there is no reason to use a Bonferroni-type procedure whose function is to control the overall error rate in multiple comparisons (Saville, 1990; Soboroff, 2014). To allow a fair comparison, a Bonferroni procedure was also not applied to the Fisher-Pitman tests.

3.1. Study 1: Song et al. (2019) on ABSA

The first case study is based on the Laptop review dataset of the SemEval 2014 Task 4 on aspect level sentiment classification (ABSA) (Pontiki et al., 2014; Song et al., 2019). This task aims at identifying the sentiment expressed towards an aspect in a sentence, using three sentiment polarities: positive, neutral and negative. The metrics to evaluate the performance was Accuracy (but some researchers reported also Macro-F1). I obtained the predictions on the challenge evaluation set by five systems using the open source ABSA-PyTorch implementation[3] provided by Song et al. (2019) and avail-

[3] ABSA-PyTorch implements more systems than the five included here, but, as this software was designed to compare their model sizes, the parameters used by the different procedures are not optimized and the results I have obtained are sometimes significantly lower than those published in the original studies. This does not pose a problem for the

able at https://github.com/songyouwei/ABSA-PyTorch (Youwei Song, 2019): AEN BERT and BERT SPC of Song et al. (2019), MemNet (Tang et al., 2016b), ATAE LSTM (Wang et al., 2016) and TD LSTM (Tang et al., 2016a).

Figure 2 shows the five systems, their accuracy on the evaluation set, the conclusions of the Fisher-Pitman test and the CIs for the differences between the systems. If the leaderboard approach is used, it is enough to look at the order of the systems; no statistical analysis is necessary. The significance tests indicate that AEN BERT and BERT SPC outperform the other three systems and that MEMNET outperforms TD LSTM.

The analysis of the CIs leads to more nuanced and richer conclusions. As an example, the difference observed, the best estimate of the true difference, between BERT SPC and MEMNET, is almost 0.05, a difference that seems to have some importance for an application. The CI at 95% for this difference indicates that it could even go as far as 0.08 accuracy, but also that it could be as low as 0.013, a much more negligible value. These observations could lead to the conclusion that BERT SPC is potentially a substantial improvement over MEMNET, but that further comparisons on independent data would be welcome, especially as MEMNET has fewer parameters. This analysis is not as favorable to the BERT models as the one conducted by Song et al. (2019), which was based only on the leaderboard, but it should be kept in mind that these authors analyzed several datasets and that the scores obtained are not perfectly identical.

On the other hand, the data collected clearly argue in favor of AEN BERT and BERT SPC compared to TD LSTM since the CI lower limit is greater than 0.045.

Figure 2 might give the impression that all the CIs are actually roughly the same length, reducing the interest of using them. If indeed four out of ten CIs have a length between 0.0689 and 0.0706, the largest length (0.0783 for BERT SPC and TD LSTM) is 125% of the smallest (0.0627 for ATAE LSTM and MEMNET). More generally, while there is a strong relationship between the conclusions of the tests (i.e, the a, b and c groupings) and those of the CIs, the crucial point is that the CIs make it possible to qualify the conclusions in relation to the simple significant vs. not significant dichotomy. For example, Figure 2 shows that there is a significant difference between MEMNET and TD LSTM and not between ATAE LSTM and TD LSTM, but the lower CI limits are (almost) identical and therefore the significant difference should be put into perspective.

3.2. Study 2: LE-PC-DNN for EmoInt

The second case study illustrates the potential usefulness of CIs when performing K-fold CV. It is based on the WASSA'17 EmoInt shared task for which systems had to estimate the emotion intensity of tweets for four emotions (anger, fear, joy and sadness) on a real-valued scale (Mohammad and Bravo-Marquez, 2017). The performance was measured by means of the Pearson correlation coefficient. In this framework, (Kulshreshtha et al., 2018) proposed *LE-PC-DNN*, an improved version of the system ranked first in the challenge. I have reproduced their study using the code they provided at https://github.com/Pranav-Goel/Neural_Emotion_Intensity_Prediction (Devang Kulshreshtha and Pranav Goel and Anil Kumar Singh, 2018). In their paper, they report an ablation analysis of the performance of their system on the evaluation set by removing one by one a parallel connected component from their LE-PC-DNN architecture.

I have reproduced this ablation analysis on the evaluation set using a K-fold CV procedure, with $K = 10$, and performed twenty different 10-fold CVs, obtained by varying the random seed. It is therefore assumed here that a researcher compares two systems on the basis of only one[4] of the many possible K-fold CVs, twenty in this case study.

To save space, only the results for Joy are shown in Table 1, but the full results are available in the downloadable materials. In three of the six comparisons, the permutation tests report statistically significant differences in all CVs and the CIs show that these differences seem indeed important, being at least 0.069.

In the other three comparisons, three to seven of the CVs gave rise to statistically significant differences according to the Fisher-Pitman permutation test, but the others did not. Depending on the CV performed, the researcher's conclusion in terms of statistical significance would therefore be different.

On the other hand, for all these comparisons, the observed difference is small (≤ 0.018) and the maximum lower limits of the CIs are extremely small (≤ 0.007). It follows that none of the CVs assessed on the basis of the CIs argues for a difference which seems sufficiently important. This analysis qualifies the conclusions of the original study, as it is far from clear that all components play an actual positive role in overall performance.

This case study shows that researchers who rely on the leaderboard or on a significance test applied to the evaluation set alone, as is very often the case, will most likely obtain unreliable conclusions. They can of course proceed as here by performing a large number

case study, but it seemed preferable to compare only systems whose performances are relatively close to the published values.

[4]There is thus no reason to apply a Bonferroni correction here, even more so because it seems difficult to justify it in the case of CVs (Bestgen, 2020b).

Condition		Correlation		p-value			Lower CL		Difference		Upper CL	
C1	C2	C1	C2	Min	Max	#sig	Min	Max	Min	Max	Min	Max
Full	¬FC	.802	.802	.0315	.9779	3	-.010	.001	-.005	.005	-.001	.011
Full	¬CNN	.802	.791	.0032	.2930	7	-.005	.007	.006	.018	.017	.030
Full	¬LE	.802	.710	.0001	.0001	20	.062	.081	.084	.107	.111	.140
¬FC	¬CNN	.802	.791	.0020	.3341	5	-.005	.007	.006	.017	.018	.029
¬FC	¬LE	.802	.710	.0001	.0001	20	.062	.083	.084	.108	.110	.140
¬CNN	¬LE	.791	.710	.0001	.0001	20	.042	.066	.069	.097	.101	.133

Table 1: Results for Joy of the twenty 10-fold CV for the ablation procedure applied to LE-PC-DNN with from left to right the names of the compared conditions, the respective correlations with the gold standard, the minimum and maximum p-value and the number of significant comparisons ($\alpha = .05$) for the Fisher-Pitman tests, and the minimum and maximum values of the lower limit of the CIs (CL), of the difference in correlation between the two conditions and of the upper limit of the CIs.

of CVs, but it will be much simpler and less resource-intensive to calculate the CIs.

4. Conclusion

The two case studies presented above indicate that it is easy to obtain a CI for the difference between two systems and that it provides useful information when interpreting the results. However, this procedure is seldom used in NLP as evidenced for example by the absence of this term in the recent *Synthesis Lectures on Human Language Technologies* by (Dror et al., 2020) on *Statistical Significance Testing for Natural Language Processing*, despite the close vicinity of these two approaches.

When developing a system or performing an ablation analysis, all the necessary data to compute the CIs are available. Computing them to compare systems proposed by different researchers is easy for the organizers of a challenge. These CIs could be calculated by default on platforms like Codalab, and on Multi-Task Benchmarks like GLUE, SuperGLUE or XTREME (Wang et al., 2018; Hu et al., 2020). By simply adding the confidence limits to each score, everyone could get a clearer picture of the superiority of one system over others.

However, calculating these CIs is much more difficult for individual researchers when they want to compare their system to the SOTA ones. This could become easier if it is recommended that, when a paper applies a system to a public dataset, the predictions for the evaluation set are made freely available to everyone, for instance in hubs of ready-to-use datasets for machine learning models (Lhoest et al., 2021).

An important characteristic of the CI approach compared to the other ways of presenting learderboard results is that, for each task, it is up to the researchers or the community to decide what difference is sufficiently important. This forces researchers to go beyond the numbers and encourages them to take other parameters into account. For instance, a difference should be seen as more or less important depending on the resources necessary to obtain the performance (Dodge et al., 2019). Likewise, CIs including zero should argue for the most interesting system according to factors other than accuracy.

5. Acknowledgements

The author is a Research Associate of the Fonds de la Recherche Scientifique (FRS-FNRS). He would like to thank the reviewers for their very constructive comments.

6. Bibliographical References

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, July. Association for Computational Linguistics.

Berrar, D. and Lozano, J. A. (2013). Significance tests or confidence intervals: which are preferable for the comparison of classifiers? *Journal of Experimental & Theoretical Artificial Intelligence*, 25(2):189–206.

Berry, K. J., Mielke, P. W., and Mielke, H. W. (2002). The Fisher-Pitman permutation test: an attractive alternative to the F test. *Psychological Reports*, 90:495–502.

Bestgen, Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing*, 29:164–170.

Bestgen, Y. (2017). Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora. *Quaderns de Filologia: Estudis Linguistics*, 22:33–56.

Bestgen, Y. (2020a). Comparing lexical bundles across corpora of different sizes: The zipfian problem. *Journal of Quantitative Linguistics*, 27(3):272–290.

Bestgen, Y. (2020b). Reproducing monolingual, multilingual and cross-lingual CEFR predictions. In *Proceedings of the 12th Conference on Language*

Resources and Evaluation (LREC 2020), pages 5597–5604, Marseille, France, May. European Language Resources Association (ELRA).

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in ASR performance evaluation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 409–412, Montreal, Canada, May.

Claridge-Chang, A. and Assam, P. N. (2016). Estimation statistics should replace significance testing. *Nature Methods*, 13:108–109, February.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49:997–1003.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. (2021). The benchmark lottery. *arXiv e-prints*.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statist. Sci.*, 11(3):189–228, 09.

Dix, A. (2020). *Statistics for HCI: Making Sense of Quantitative Data*. Synthesis Lectures on Human-Centered Informatics. Morgan & Claypool Publishers.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. (2019). Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November. Association for Computational Linguistics.

Dror, R., Peled-Cohen, L., Shlomov, S., and Reichart, R. (2020). Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13:1–116.

Graham, Y., Mathur, N., and Baldwin, T. (2014). Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Hayden, R. W. (2019). Questionable claims for simple versions of the bootstrap. *Journal of Statistics Education*, 27(3):208–215.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., and Epstein, R. (2003). Bootstrap methods and permutation tests. In D. S. Moore et al., editors, *Introduction to the Practice of Statistics*, pages 1–74. W H Freeman.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III et al., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul.

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with p-values? *The American Statistician*, 73(sup1):20–25.

Ismay, C. and Kim, A. Y. (2019). *Statistical Inference via Data Science: A ModernDive into R and the Tidyverse*. CRC Press.

Kim, J. H. and Robinson, A. P. (2019). Interval-based hypothesis testing and its applications to economics and finance. *Econometrics*, 7(2):21, May.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Kulshreshtha, D., Goel, P., and Kumar Singh, A. (2018). How emotional are you? Neural architectures for emotion intensity prediction in microblogs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2914–2926, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. (2021). Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Liu, Z. (2019). Statistical testing on ASR performance via blockwise bootstrap. *CoRR*, abs/1912.09508.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2018). Are GANs created equal? a large-scale study. In S. Bengio, et al., editors, *Advances in Neural Information Processing Systems 31*, pages 700–709. Curran Associates, Inc.

Masson, M. E. J. and Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, 57:203–220.

Mohammad, S. and Bravo-Marquez, F. (2017). WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, September. Association for Computational Linguistics.

Neuhauser, M. and Manly, B. F. J. (2004). The Fisher-Pitman permutation test when testing for differences in mean and variance. *Psychological Reports*, 94:189–194.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Rao, C. R. and Lovric, M. M. (2016). Testing point null hypothesis of a normal mean and the truth: 21st century perspective. *Journal of Modern Applied Statistical Methods*, 15:2–21.

Saville, D. J. (1990). Multiple comparison procedures: the practical solution. *The American Statistician*, 44:174–180.

Schmidt, F. L. and Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, et al., editors, *What if there were no significance tests?*, pages 37–64. Erlbaum.

Soboroff, I. (2014). Computing confidence intervals for common IR measures. In *Proceedings of the Sixth International Workshop on Evaluating Information Access (EVIA 2014)*, pages 25–28.

Song, Y., Wang, J., Jiang, T., Liu, Z., and Rao, Y. (2019). Attentional encoder network for targeted sentiment classification. *arXiv e-prints*, page arXiv:1902.09314, February.

Tang, D., Qin, B., Feng, X., and Liu, T. (2016a). Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Tang, D., Qin, B., and Liu, T. (2016b). Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas, November. Association for Computational Linguistics.

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician*, 73(sup1):1–19.

Zhang, Y. and Vogel, S. (2004). Measuring confidence intervals for the machine translation evaluation metrics. In *In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2004*, pages 4–6.

7. Language Resource References

Youwei Song. (2019). *Aspect Based Sentiment Analysis, PyTorch Implementations*. available at https://github.com/songyouwei/ABSA-PyTorch.

Alexander Neshitov. (2019). *A small package for bootstrap confidence intervals*. available at https://pypi.org/project/bootstrap-interval/.

Devang Kulshreshtha and Pranav Goel and Anil Kumar Singh. (2018). *How Emotional Are You? Neural Architectures for Emotion Intensity Prediction in Microblogs*. available at https://github.com/Pranav-Goel/Neural_Emotion_Intensity_Prediction.

Yves Bestgen. (2022). *BootCI: Bootstrap confidence intervals and Fisher-Pitman test for paired samples*. available at https://pypi.org/project/BootCI/.