

# It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks

Valerio Basile<sup>a</sup>

<sup>a</sup>University of Turin, Corso Svizzera 185, Turin, Italy

## Abstract

Supervised machine learning, in particular in Natural Language Processing, is based on the creation of high-quality gold standard datasets for training and benchmarking. The de-facto standard annotation methodologies work well for traditionally relevant tasks in Computational Linguistics. However, critical issues are surfacing when applying old techniques to the study of highly subjective phenomena such as irony and sarcasm, or abusive and offensive language. This paper calls for a paradigm shift, away from monolithic, majority-aggregated gold standards, and towards an inclusive framework that preserves the personal opinions and culturally-driven perspectives of the annotators. New training sets and supervised machine learning techniques will have to be adapted in order to create fair, inclusive, and ultimately more informed models of subjective semantic and pragmatic phenomena. The arguments are backed by a synthetic experiment showing the lack of correlation between the difficulty of an annotation task, its degree of subjectivity, and the quality of the predictions of a supervised classifier trained on the resulting data.

## Keywords

Linguistic Annotation, Subjectivity, Inclusive Machine Learning

## 1. Introduction

Much of modern Natural Language Processing (NLP) and other areas of Artificial Intelligence (AI) are based on some form of supervised learning. In the past decades, models like Hidden Markov Models, Support Vector Machines, Convolutional and Recurrent Neural Networks, and more recently Transformers had represented the state of the art in many NLP tasks. However different the architectures may be, the common basis of supervised statistical models is data produced by humans by some process of *annotation*.

Linguistic annotation has always been a staple of the creation of language resources, which are employed as training material for supervised models as well as for benchmarking and to compare the performance of systems. The annotation for a language resource is a pretty standardized process. The techniques involved in the process come from the linguistic tradition and have been incorporated into the toolkit of the modern computational linguist. Such techniques include annotation by multiple subjects, measures of inter-annotator agreement, harmonization,

---

*AIXIA 2020 Discussion Papers Workshop*

✉ [valerio.basile@unito.it](mailto:valerio.basile@unito.it) (V. Basile)

🆔 0000-0001-8110-6832 (V. Basile)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

aggregation by majority, and so on.

In parallel to the evolution of more and more technologically advanced statistical models, the focus of the attention of the NLP community has also shifted from more “low level” linguistic phenomena such as part-of-speech tagging and syntactic parsing, to more and more “high level”<sup>1</sup> tasks depending on extra-contextual cues and world knowledge. Seen from another angle, in recent years the attention has grown towards more and more **subjective** tasks such as sentiment analysis, irony detection, up to highly subjective tasks such as hate speech detection.

In this paper, I will highlight the main issues that arise when applying traditional language annotation methodologies to highly subjective phenomena. Starting with a brief reminder on the basic principles of standard annotation procedures, I will show how a paradigm shift is needed in order to fully model complex, multi-perspective language phenomena. I will then propose new directions to follow in order to foster the development of a new generation of inclusive supervised models, presenting the results of a simulated experiment, as well as evidence from recent literature, to support the claims.

## 2. A Quick Primer on Linguistic Annotation

To prepare the ground, let us introduce the basic principles of the process of manual annotating linguistic data. The main components of an annotation task are the following:

- A set of **instances** to annotate. These can be sentences, documents, words (in or out of context), or other linguistically meaningful units.
- A target **phenomenon**, described in detail by means of guidelines and examples.
- An annotation **scheme**, defining the possible values for the phenomenon to annotate, and additional rules, where applicable.
- A group of **annotators**, selected on the basis of expertise, availability, or a mix of the two.

The annotation process is an iterative process, where each **annotator** expresses their judgment about the target **phenomenon** on one **instance** at a time, in the modalities defined by the annotation **scheme**. The possible values may be categorical variables, real numbers, integers on a scale, and so on.

The annotation is usually carried out by either experts and the crowd. Experts are a broad category comprising people considered competent on the phenomenon that is being annotated. However, this category has grown to include people that are not necessarily experts in certain phenomena by academic standards, but rather they present characteristics deemed relevant to a specific annotation, such as, for instance, victims of hate speech, or activists for social rights, in abusive language annotations [1]. Finally, experts are often simply the authors of the work involving the annotation, their associates, students, or friends. That is, expert annotation is often times a matter of availability of human resources to perform the annotation task.

---

<sup>1</sup>The metaphor refers to the ideal spectrum often used in linguistics, where phenomena of natural language are organized on a scale roughly covering, in order: morphology, syntax, semantics, pragmatics.

Since the annotation of language data is notoriously costly, in the last decade scholars have turned more and more to crowdsourcing platforms, like Amazon Mechanical Turk<sup>2</sup> or Appen<sup>3</sup>. Through these online platforms, a large number of annotators are available for a reasonable price.<sup>4</sup> The trade-off, when using these services, is a lesser control on the identity of the annotators, although some filters based on geography and skill can be imposed. Moreover, as the number of annotators grows, the set of instances to annotate is divided among them unpredictably, and the participation of each individual to the annotation task is typically uneven. As a result, with crowdsourcing, the question-answer matrix is sparse, while it is in general complete with expert annotation.

Once a sufficient number of annotations on a sufficient number of instances is collected, they are compiled into a *gold standard* dataset that represents the truth against which comparing future predictions on the same set of instances, much like the gold standard in financial terms it gets its name from<sup>5</sup>. The most straightforward procedure to compile a gold standard from a set of annotations is to apply some form of instance-wise aggregation, such as by majority vote: for each instance, the choice indicated by the relative majority of the annotators is selected as the true value for the gold standard. Depending on a series of factors including the number of annotators, this phase can be more or less complicated, e.g. involving strategies to break the ties, or compute averages in the case of the annotation of numeric values on a scale. Sometimes, extra effort is put into resolving the disagreement. This is done by thoroughly discussing each disagreed-upon instance, going back to the annotation guidelines, or having an additional annotator make their judgment independently, or any combination of these methods. This phase takes the name of *harmonization*.

Quantitative measures of inter-annotator agreement are computed to track how much the annotators gave similar answers to the same questions. Among the most popular ones we find percent agreement (the ratio of the number of universally agreed-upon instances over the total number of instances), Cohen's Kappa (a metric that takes into account the probability of agreeing by chance), Fleiss' Kappa (a generalization of Cohen's Kappa to an arbitrary number of annotators), and Krippendorff's Alpha (a further generalization applicable to incomplete question-answer matrices). One of the purposes of computing inter-rater agreement is to provide a quantitative measure of how hard the task is for the human annotator. As such, the inter-annotator agreement is also interpreted as related to the upper bound of measurable computer performance on the same task. The inter-annotator agreement is typically computed before harmonization, sometimes both before and after, in order to measure the efficacy of the harmonization itself.

Lately, techniques from the Content Analysis community are being more and more integrated into the annotation process for machine learning purposes. Among these, it is not unusual that a small sample of instances are annotated by all the available annotators and the inter-annotator agreement metrics are computed on this set. The small sample is often called *test set*, which should not be confused with the meaning of the same term in machine learning lingo (a set of instances used to test the performance of a model). After the small sample is annotated,

---

<sup>2</sup><https://www.mturk.com/>

<sup>3</sup><https://appen.com/>

<sup>4</sup>Whether this price is fair has been debated for some years now [2]

<sup>5</sup>[https://en.wikipedia.org/wiki/Gold\\_standard](https://en.wikipedia.org/wiki/Gold_standard)

if the computed agreement is found satisfactory (e.g., above a predetermined threshold), the annotation continues by splitting the rest of the dataset among the annotators, who proceed independently from one another. While this methodology is capable of producing large amount of annotated data in shorter time, which is important especially in the era of deep learning, it does not solve the other issues which I raise in the rest of this paper.

### 3. The Annotation of Highly Subjective Phenomena

In this article, I am referring to a “subjective” task in the sense of a linguistic tasks for which the human judgment is inherently influenced by factors pertaining to the judges themselves, rather than the linguistic phenomenon. In contrast, human judgment on an “objective” task depends uniquely on the object to be judged. As a corollary, different judgments on an objective task should ideally always coincide, barring negligible amounts of measurement noise, while the same does not apply to subjective tasks.

One of the aims of this paper is to stimulate the discussion on the subjectivity of NLP tasks, how it affects their evaluation, and, ultimately, the development of systems capable of solving them. On an ideal scale from total objectivity to total subjectivity, traditional tasks in Computational Linguistics such as part-of-speech tagging sit towards the former end. During a POS-tag annotation, inconsistencies can be found among the annotations coming from different judges. However, these are typically caused by a different interpretation of the rules, or genuine mistakes, rather than actual, heartfelt disagreement or divergence of opinions. On the contrary, while annotating a highly subjective tasks such as offensive language, different people could find different expressions offensive to very different extent. I argue that in such cases, **all the opinions** of the annotators **are correct**.

**Proposition 1.** *Disagreeing annotations that comes from diverging opinions should be equally considered in the construction of a gold standard dataset.*

Unfortunately, traditional annotation methodologies do not leave space to implement such proposition. The reason is that language annotation operates under the unwritten postulate that *there is exactly one truth*, i.e., the correct annotation towards which human judgments converge. Multiple annotations and aggregation by majority are the main tools to facilitate this convergence. However, **in the subjective task scenario, the one-truth assumption does not hold** anymore.

In standard linguistic annotation, agreement metrics are used to measure the difficulty of a task and the common understanding of the annotation guidelines by the annotators. Applied to a subjective task, agreement metrics inevitably capture divergence of opinions as well, mixing the signals into a single quantitative measure that therefore loses its meaning to a certain extent. To be fair, issues with current agreement metrics have been highlighted in recent literature [3]. Alternative metrics have been proposed that take disagreement into account [4], and frameworks to leverage the informative content of annotator disagreement have been implemented [5, 6]. Some approaches address issues with the annotation methodology by tackling annotator reliability [7]. Perhaps the work that is most in line with the position expressed in the present paper is [8], which shows by statistical tests how "harmonization

sometimes harms", and propose to use a weighting scheme based on individual annotations to improve the evaluation of NLP models for subjective tasks. In a recent paper, we propose a stronger version of such idea, in order to account for all the perspectives of a set of annotators, extracting the automatically and weighting them equally [9].

To address the issues described so far, we argue for two positions, complementary to each other. The first is a position **against the release of aggregated datasets** for benchmarking AI (and NLP in particular) models. The second is a position **for a new evaluation paradigm** for highly subjective NLP tasks, that takes multiple perspectives into account. These positions are detailed in the next sections.

## 4. The Power of Pre-aggregated Data

In our own previous work, we have shown how leveraging the divergence of opinions of the human annotators of particularly subjective tasks can lead to an improvement of the quality of the annotated dataset for training purposes [10]. In that work, we defined a quantitative index to measure the *polarization* of the judgments on single instances as a distinct concept from inter-annotator agreement. Specifically, we employed the polarization index to filter out instances from hate speech detection benchmark datasets that showed a high degree of polarization, and give more weight to the less polarizing instances. The training set resulting from this transformation was found to induce a better model for the hate speech detection task, indicating that indeed the high subjectivity of the phenomenon tends to confuse the supervised classifier.

In a subsequent work [9], we took this approach one step further, by training separate classifiers to model different, automatically extracted perspectives of the annotators on the same instances. We trained an “inclusive” classifier that takes into account all of the extracted opinions, including the ones expressed by a minority of the annotators. Such inclusive classifier proved to work better than all the others in the highly subjective task of hate speech detection.

The common denominator of these works is that these approaches need access to the *pre-aggregated* annotated data, i.e., every single annotation on the instances of the training dataset. The lesson learned is that the fine-grained information contained in the pre-aggregated, complete annotation is extremely valuable in order to model different perspectives on a linguistic phenomenon, with particular importance towards subjective phenomena. Therefore, I put forward a call to action for every NLP researcher and developer of language resources:

**Proposition 2.** *Manually annotated language resources should always be published along with all their single annotations.*

## 5. Perspective-aware Evaluation

The problem of modeling the personal point of view of the annotators, however, is only partially solved by the approach presented so far. While a perspective-aware model can fare well on a standard benchmark, if the test set is constructed by means of aggregation (e.g., by majority voting on each instance) the evaluation will not be fair with respect to the multiple perspectives.

In other words, building a system capable of encoding multiple perspectives (by leveraging the information in a pre-aggregated dataset) is of limited use if such perspectives are not represented in the testing benchmark. The model would still be forced to produce one single label (or any other kind of single output) in order to match it with the gold standard test set. On the contrary, a benchmark where test sets are themselves in a pre-aggregated form would enable a complete and fair evaluation with respect to all the perspective encoded in such test data. I therefore propose to radically change the way we test NLP systems, by taking into account the diverging opinions of the annotators throughout the entire evaluation pipeline:

**Proposition 3.** *Predictive models for highly subjective phenomena should be tested against pre-aggregated benchmarks.*

The problem remains open of what kind of evaluation metrics one can use to carry out such perspective-aware evaluation. In the next section, I present an experiment with synthetic data, and showcase one possible methodology of evaluation, showing how it is effective, to a certain extent, at separating the quantitative measurement of the difficulty of a NLP task from its subjectivity.

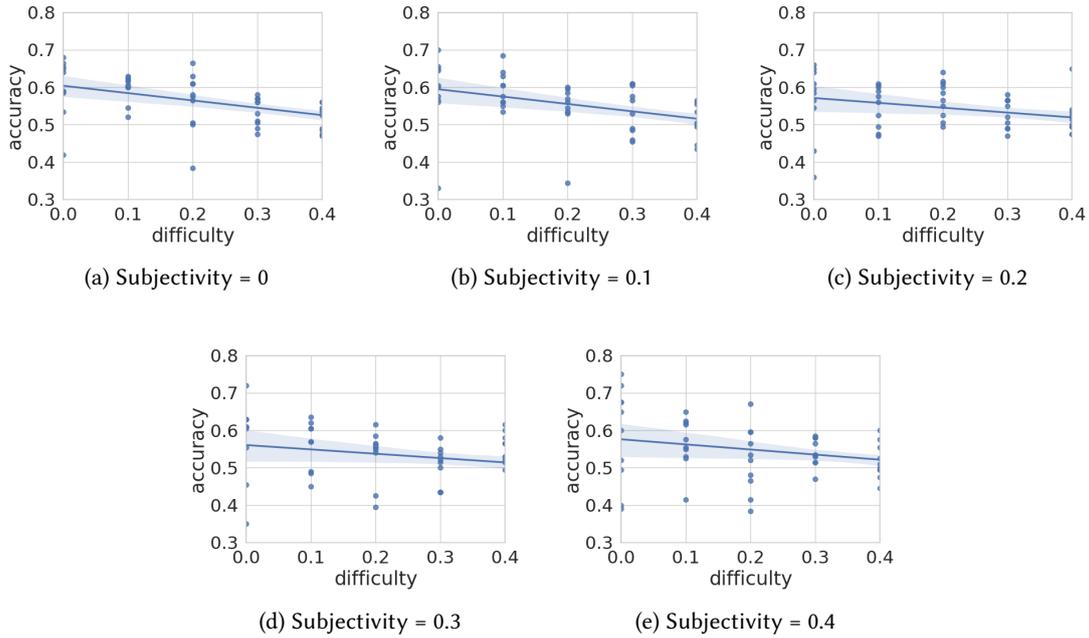
## 6. An Experiment with Synthetic Data

In this section, an experiment is shown to further drive the points argued in the paper so far. The experiment is a simulation on synthetic data, presented in an attempt to exemplify the main arguments of this proposal with no additional real-world noise, rather than to show the practical effectiveness of a method implementing those principles.

The simulation involves an annotation task, with 10 annotators and 1,000 instances. The task is a binary classification, whereas the annotators are asked to mark each instance as either 0 or 1 (or true/false, black/white, or any other binary distinction). Each instance is encoded as a series of 100 binary features. The annotators have a “background”, i.e., they are equally split into two groups.

Two parameters are set that influence the annotation, namely difficulty and subjectivity. A higher difficulty means that an annotator has a high chance of labeling an instance with the wrong label. Subjectivity is more subtle and interplays with the annotators’ background. For each instance, there is a chance (depending on the value of the subjectivity parameter) to be a “subjective” instance. If that is the case, the label will depend on the background of the annotator, unless a wrong annotation is given because of the difficulty of the task. Finally, the features are computed to correlate with the annotations, with 20% random noise artificially injected. The expected accuracy of a cross-validation experiment on this dataset, with zero difficulty and zero subjectivity in the annotation process, is around 80%.

The simulation is run ten times for each combination of the values of the two parameters in the range 0–0.4 in 0.1 steps, each run producing a full set of annotations, and a gold standard aggregated by majority voting. Each of these datasets is used in a 10-fold cross-validation supervised learning experiment, to assess the quality of the annotation in a standard machine learning scenario. The classifier is a deep multi-layer perceptron with two 10-node hidden layers and a single output node. Nodes at the hidden and output level are equipped with a sigmoid



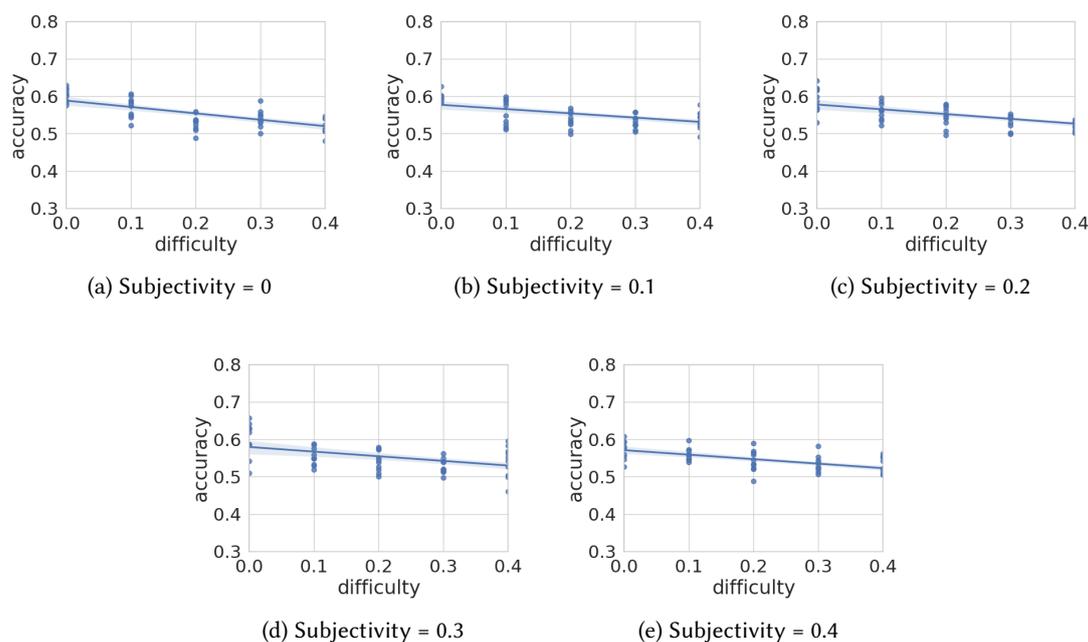
**Figure 1:** Correlation between difficulty of the annotation task and the accuracy of a classifier trained on the resulting dataset in a cross-validation experiment.

activation function. For the purposes of this experiment, variation in the size of the network, activation function (e.g., sigmoid vs. linear), and hyperparameters were not critical, in that they did not change the conclusions in any significant way. The result of the cross-validation is a single figure for accuracy. We plot it, repeated for ten runs for each value of the hyperparameter space (difficulty and subjectivity of the annotation task) in multi-plots in Figure 1.

The plots show the expected negative correlation between accuracy and difficulty. It is not surprising that a difficult task will produce a dataset that is less informative to a supervised model, resulting in worse performance in cross validation. However, comparing the plots across increasing values for subjectivity, the correlation becomes less accentuated. The more a task is subjective, the less the evaluation is correlated with its difficulty alone. This is an indication that subjectivity and difficulty are indeed different phenomena, while standard evaluation methodology measures their respective signals at the same time.

The same experimental setting can also be used to test whether another evaluation framework is feasible, where aggregated data are avoided altogether. Here, the experiment is run exactly like in the previous iteration, except that separate classifiers are trained on each column of pre-aggregated labels individually, and tested accordingly. The final accuracy score is simply the arithmetic mean of the ten annotator-specific accuracy scores. The results of this second experiment are shown in Figure 2.

These plots, compared to the previous set, show an interesting pattern. The negative correlation between difficulty and accuracy is much clearer when the evaluation is done on pre-aggregated data, as shown by the much narrower intervals where points lie on the y-axes.



**Figure 2:** Correlation between difficulty of the annotation task and the accuracy of a classifier trained on the resulting dataset in a cross-validation experiment.

This is to be interpreted as evidence that indeed all the opinions from the annotators matter, not only in principle, but also towards a more fair evaluation for classifiers of subjective phenomena.

## 7. Conclusion

In this paper, I argued for a paradigm shift regarding how language resources are created, published, and incorporated into experimental pipelines for benchmarking. I have shown how the methodology for manual annotation generally employed to create language resources, which comes from the linguistic tradition, suffers from a new set of issues when it is applied to NLP tasks that are becoming more prominent in recent times, focusing in particular on the problem of subjective tasks.

Following the development of recent literature, I formulated two recommendations, in an effort to stir the discussion about what I consider critical problems to solve for the next generation of NLP systems, and the future of a perspective-aware AI. To further drive the point across, I proposed an experiment with simulated data, to highlight *in vitro* what is the impact of my proposal on real world evaluation procedures.

To be fair, the international Natural Language Processing Community is starting to be sensitive to these ideas. An example is the shared task 12 organized at SemEval 2021 on Learning with Disagreements, where six datasets are proposed to the participants in their pre-aggregated form.

As a conclusive remark, the thoughts expressed in this paper are, in a way, a formalization of a series of reflections coming from the author’s experience and, to a great extent, feedback from and discussion with a number of scholars sensitive to the issues I raised here. As such, I believe the AI community is already mature to accept the next step towards perspective-aware models and to recognize that more than one truth is possible when perception plays an important role in language-mediated communication. This work represents therefore just one possible way to implement such change.

## Acknowledgments

The author would like to express his gratitude to the anonymous reviewers, whose comments greatly contributed to improving this work for the final version. This work is partially funded by the project “Be Positive!” (under the 2019 “Google.org Impact Challenge on Safety” call).<sup>6</sup>

## References

- [1] Z. Waseem, Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 138–142. URL: <https://www.aclweb.org/anthology/W16-5618>. doi:10.18653/v1/W16-5618.
- [2] A. Felstiner, Working the crowd: employment and labor law in the crowdsourcing industry, Berkeley J. Emp. & Lab. L. 32 (2011) 143.
- [3] D. M. W. Powers, The problem with kappa, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 345–355. URL: <https://www.aclweb.org/anthology/E12-1035>.
- [4] A. Checco, K. Roitero, E. Maddalena, S. Mizzaro, G. Demartini, Let’s agree to disagree: Fixing agreement measures for crowdsourcing, in: Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada., AAAI Press, 2017, pp. 11–20.
- [5] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, AI Magazine 36 (2015) 15–24.
- [6] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A case for soft loss functions, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (????) 173–177. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/7478>.
- [7] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, E. Hovy, Learning whom to trust with MACE, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 1120–1130. URL: <https://www.aclweb.org/anthology/N13-1132>.
- [8] M. Klenner, A. Göhring, M. Amsler, Harmonization sometimes harms, in: S. Ebling, D. Tuggener, M. Hürlimann, M. Cieliebak, M. Volk (Eds.), Proceedings of the 5th Swiss

---

<sup>6</sup><https://impactchallenge.withgoogle.com/safety2019>

Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only], volume 2624 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: <http://ceur-ws.org/Vol-2624/paper10.pdf>.

- [9] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020)* 151–154. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/7473>.
- [10] S. Akhtar, V. Basile, V. Patti, A new measure of polarization in the annotation of hate speech, in: M. Alviano, G. Greco, F. Scarcello (Eds.), *AI\*IA 2019 - Advances in Artificial Intelligence - XVIIIth International Conference of the Italian Association for Artificial Intelligence*, Rende, Italy, November 19-22, 2019, *Proceedings*, volume 11946 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 588–603. URL: [https://doi.org/10.1007/978-3-030-35166-3\\_41](https://doi.org/10.1007/978-3-030-35166-3_41). doi:10.1007/978-3-030-35166-3\_41.