Towards Inter-Rater-Agreement-Learning

Kai-Jannis Hanke, Andy Ludwig, Dirk Labudde and Michael Spranger
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*name.surname*}@hs-mittweida.de

***Abstract*—While technological advances and improved algorithms enhance most scientific fields, there remains a simple problem in many domains. If a decision has to be made we resort to simple majority votes or utilize agreement measures to determine how unanimous a decision is. Especially in text classification, a text is usually sorted into a specific category based on how many people agreed on it. However, the problem is that in these methods the individual that made the decision is neglected. Therefore, we propose a weighted approach that includes a flexible feature space and adjustments to the weights not only according to the individual's expertise but also to their performance on previous tasks. Preliminary experiments with a data set including short music related texts yield promising results with fewer cases for which no majority vote was achieved.**

Keywords–Agreement Measures; Weighted Majority Vote; Text Labeling.

I. INTRODUCTION

Scientific inquiries require sound and accurate measurements to ensure valid and reliable test results. While machine learning and artificial intelligence are becoming more prevalent in different research applications, the final judgment or the required previous labeling of data mostly remains a group decision of trained or untrained individuals. Human judgment is often considered gold standard. May it be text annotation for subsequent text classification, diagnoses by a collective team of medical professionals [1] or generally phrased: the correct labeling or interpretation of data in any other domain. This gold standard is frequently acquired by utilizing a majority decision.

However, ordinary majority votes neglect the unique strengths and weaknesses of the individuals participating in a given labeling experiment. In the case of a complicated medical case, a trained medical expert with decades of experience should have more influence on the final judgment than a student on their first internship. Obviously, the expertise and experience of an individual should play a decisive role, especially when it comes to human life. Still, in less lethal decisions the competence of individuals is often not taken into account. Instead, each individual receives equal importance in the final judgment, regardless of their predisposition to the topic at hand. Decisions of individuals with a keen interest in artificial intelligence and machine learning should have more weight when labeling texts about aforementioned topics than a nutritionist. The expertise on a topic should always be taken into consideration if the desired outcome is a gold standard. Furthermore, if an individual has the desire to perform well on a given task then this should induce a positive weight, whereas the goal of finishing as quickly as possible should lead to decreased influence in a majority vote. Additionally, internal consistency or the certainty with which individuals perform may also provide insight into the value of their contribution. If presented with the same task at different time points, then ideally the answers should be the same. If however there is a discrepancy, i. e,. if the individual makes a different judgment when repeatedly coming across the same question, then their overall value for the given assignment diminishes. In order to receive an ideal gold standard, these problems need to be addressed and implemented.

Hence, in this paper we propose a weighted majority vote to derive an increased amount of gold standard labels compared to an unweighted majority vote. When making majority decisions we inevitably come across items with no majority, but by implementing a flexible feature space we are able to push the boundary a little further and receive majorities, based on expertise, response time and internal consistency even when an unweighted approach did not meet the majority criterion.

This approach will be used to annotate a multilingual data set with music related texts in order to create a gold standard for the development of a cross-language classification algorithm. First steps were taken in testing the approach using an already labeled single-language data set from the music domain. However, for these preliminary experiments it was only possible to use those features that were already available when the data set was labeled.

The next section provides an overview of related work, whereas Section III explains the features we used for our weighted approach. Afterwards, the used approach is explained in Section IV. The results of the preliminary experiments are presented in V and the conclusion and future work in VI.

II. RELATED WORK

The quality of the data used for training text mining systems has significant influence on the final results. Thus, enhancing data quality is a requirement, if we want to receive improved outcomes. In case of processing data labeled by human beings, various quality altering approaches are known. One way of determining the integrity of a data set is to measure the extent to which humans agree on its content. The process of evaluating accordance is called Inter-Rater-Reliability or Inter-Agreement. Furthermore, the procedure for annotating texts and finding the agreement between annotators is also called Inter-Rater-Agreement-Annotator-Agreement.

An overview about methods measuring agreement among corpus annotators is given in a survey article by Artstein and Poesio [2].

The fundamental idea to assess the quality of ratings is to measure the agreement among individuals. A simple approach is the percentage of agreement as described by Scott in 1955 [3]. A combination of this observed agreement with an expected agreement is called chance-corrected agreement. The most common and widely used agreement approaches are

Cohen's kappa coefficient [4] for two raters and Fleiss' kappa coefficient [5] for more than two raters. An advanced approach using calculated weights is given by Cohen in 1968 [6].

A different method is to calculate the agreement by evaluating the disagreement, for example in Krippendorff's alpha coefficient [7].

Those measures are based purely on statistical analyses of the given data, they do not integrate additional features. Individual biases have to be considered to gain better results. For example the emotional state of a rater influences their judgment when dealing with emotional content [8]. Thus, individual stress levels or state of mind could be used as a feature to adjust individual weights in the final decision.

Another important point is the problem of global agreement, meaning the overall agreement of all raters. Usually, methods calculate global agreement. However, this procedure may mask the actual complexity and heterogeneity of the given data [9]. Hence, looking at a type of local agreement considering subgroups of raters with similar properties, for example, experts in a certain area, might be a starting point for accurately representing a data corpus. Moreover, Boguslav et. al indicated that the agreement between annotators might not actually be the upper bound for machine learning tasks [10]. A human labeled data set for which statistical agreement measures were calculated is not necessarily the gold standard but rather a heuristic, since there might be algorithms delivering better results than the human counterpart.

Lastly, there are major problems with basic agreement statistics. Inherently, agreement does not necessarily reflect facts. If an item belongs to topic x and two raters label it as y while only one expert rater judges it to be x, then the agreement is still in favor of y even though the actual topic is x. Furthermore, chance adjusted agreement has distinct problems in both directions. If we have low chance agreement, the influence after adjusting is marginal, thus chance adjusted agreement merely becomes ordinary agreement. On the other hand, high chance agreement can yield low chance adjusted agreement even when the individual raters have good accuracy. Hence, Passonneau and Carpenter [11] show the advantages of using the Dawid Skene model [12], which leads to avoiding problems of Cohens's kappa statistic.

III. The Ideal Feature Space

Usually, text labeling is based on a connection of individual decisions of a group of individuals. However, this connection reduces the whole decision making process to one single value whereas an important aspect is lost, the individual. Hence, we included additional features that can either be measured during or collected previous to the labeling process to receive a more holistic picture. Namely, these include the response time per rater and item, the internal consistency measured by Intra-Rater-Agreement, the conscientiousness per rater, the language proficiency and, finally, the topic expertise. For example, a native speaker should be given more value than a non-native speaker and in the same manner an expert on a given topic should have more weight than an individual barely having any knowledge in that subject area. In order to see how reliable a given individual is, we calculated their Intra-Rater-Agreement. It tests whether a rater expresses the same decision for the same item on different occasions. If decisions are frequently followed by uncertainty, then choices of this annotator cannot be valued as highly as the decisions from someone who is at least partially certain.

Furthermore, previous to a labeling process a questionnaire can be used to probe the conscientiousness of an individual rater to see how reliable and trusting this individual would judge themselves, even though self-judgment can be a two-edged sword as individuals might see themselves in a biased way, as John and Robins have shown [13]. Conscientiousness can be measured by using questions from a personality test using the "Big Five" personality dimensions as introduced by [14], which is widely accepted and has been used on many occasions in psychological studies. For example, a test based on the five personality dimensions was used to link personality traits to job performance as shown by Barrick and Mount [15] but also to overall career success [16]. As the multilingual data set for the planned gold standard is going to be labeled by German speaking annotators, it is planned to use the German personality test using the five personality dimensions as introduced by Satow [17]. Satow's questionnaire contains ten questions probing the conscientiousness of an individual and contrary to the classical "Big Five", Satow's scale goes from one to four and not to five. The reason behind this is to avoid an inherent tendency towards the middle. While we do not want to force a labeling decision on inconclusive texts in our data set, we do ideally push raters to take a more extreme stance on their self judged conscientiousness to ensure diversity in our data set and to avoid leaning towards the middle. Language proficency can be measured, for example, on a scale of one to six representing the different levels (A1-C2) of the "Common European Framework of Reference for Languages" (CEFR) [18] and later normalized to a scale of 1 to 5 to fit in with the other scales. Subsequently, the topic expertise needs to be judged by the individual raters themselves and can be measured on a scale from one to five. Finally, we also want to reward raters who perform well on multiple items. Thus, if a rater's decision for a specific item is in agreement with the majority of all raters for this item they shall be rewarded in subsequent decisions and, otherwise, be punished. As a result, experts for the labeling process might emerge that were not imminent by merely looking at the collected personal information.

IV. Weighted Learning Approach

Let J be a set of raters and I a set of items that are being labeled by these raters. In a classic majority vote, each rater $j \in J$ has the same weight $w = 1$ on every item $i \in I$, regardless of their item specific competence. The desired outcome is to adjust the original weight w for each individual rater j and for the individual items i, such that w_{ji} varies from item to item and from rater to rater depending upon the feature values that have been discussed in the previous section. The weight for rater j and an item i is written as w_{ji} and calculated in (1).

$$w_{ji} = R_j - f(t_{ji}) \tag{1}$$

R_j denotes the rater's competence and is a combination of Intra-Rater-Agreement, topic expertise and language proficiency. Further, $f(t_{ji})$ takes specific values depending upon the response time and conscientiousness for a given rater. In this case, t_{ji} refers to the time an individual j needed to label a given item i.

For the planned annotation of the multilingual data set we consider to include two measures for the topic expertise. Firstly, how frequently a person was in contact with music, e.g., listening to music and, secondly, how regularly they attended events with a high emphasis on music, such as festivals, parties or concerts.

As seen in (2) R_j utilizes n different features x. For the planned gold standard we will consider $x = 4$ features, namely the self-judged music event attendance, overall contact with music, language proficiency and Intra-Rater-Agreement. Additionally, each feature has a unique weighing parameter β which not only enables prioritizing certain features, but also optimizing the weights to receive ideal results.

$$R_j = \frac{\sum_{l=1}^{n} \beta_l x_{lj}}{\frac{1}{|J|}\sum_{j'=1}^{|J|}\sum_{l=1}^{n} \beta_l x_{lj'}} \tag{2}$$

The simplistic idea behind the rater competence R_j is to reward individuals that perform better than the average rater on the sum of features.

Furthermore, for the function $f(t_{j1})$ combining response time and conscientiousness-scores we need to bring the two in a similar format. We achieve this by $[0, 1]$ normalizing the conscientiousness-scores C for each rater j and the response times t for the individual items i, as normalizing over the entire set of items would mitigate the text length of the individual items. In the weighing process the relation of response time t and conscientiousness-score C will give each rater a unique interval. The baseline is the average response time $\bar{t_i}$ for a given item. If rater j has a bigger conscientiousness-score than the normalized average response time then their interval is $[\bar{t_i}, C_j]$. When the response time for the specific item t_{ji} falls within the specific interal the labeler will not face consequences, falling farther away from the given interval results in an increasing penalization. The same procedure applies to a conscientiousness-score lower than the average $[C_j, \bar{t_i}]$. The general function $f(t_{ji})$ is seen in (3) and it helps in pointing out individuals that may not be entirely focused on the task at hand. It can be assumed that individuals with low conscientiousness finish labeling tasks rather quickly while high conscientiousness individuals may need more time to come up with a decision. Contrary, if a low conscientiousness individual takes unusually long it could be due to distractions or lack of focus, which results in a slight penalization. Analogously, individuals with high conscientiousness just rushing through the labeling decisions are penalized, since we expect them to take more time before reaching a decision.

$$f(t_{ji}) = \begin{cases} 0, & C_j > \bar{t_i} \land t_{ji} \in [\bar{t_i}, C_j] \\ 0, & C_j < \bar{t_i} \land t_{ji} \in [C_j, \bar{t_i}] \\ t_{ji} - \bar{t_i}, & C_j > \bar{t_i} \land t_{ji} \notin [\bar{t_i}, C_j] \\ \bar{t_i} - t_{ji}, & C_j < \bar{t_i} \land t_{ji} \notin [C_j, \bar{t_i}] \end{cases} \tag{3}$$

For subsequent weights $w_{j(i+1)}$ how well rater j performed on the previous decisions is included. In this context performing well means being part of the majority for the previously weighted decision according to a pre-defined rule that defines majority. If rater j performed excellently, then part of their previous weight $w_{(i-1)j}$ will be transferred to the current decision w_{ij}, allowing them an increased error margin. At the same time, we do not want to over-penalize individuals that underperformed on previous items. Therefore, a parameter b is introduced to regulate how much of the previous weight can be used for the current item. In (4) the parameter b is defined. If an individual performed perfectly, e.g., is part of the majority for every single decision, then $b = 1$. Drastic underperformance yields a b approaching 0.

$$b_{ji} = \frac{1}{\lambda(i-1)} \sum \begin{cases} 1, \text{for j in majority for item i} \\ 0, \text{for j in minority for item i} \end{cases} \tag{4}$$

Since $b = 1$ results in an overweighting of the previous decision and completely neglects the current decision, we also utilize the parameter λ to declare an upper bound for b. Thus, $\lambda = 1$ would result in $b = 1$, whereas $\lambda = 2$ delivers a more reasonable $b = 0.5$. This has the advantage of giving experts some flexibility to bring in their previous expertise and turn the vote in their favor whereas underperforming raters can still utilize their estimated expertise for the current item without being too heavily penalized for their previous performance.

$$w_{ji} = (1 - b_{ji})[R_j - f(t_{j(i+1)})] + \frac{b_{ji}}{i} \sum w_{j(i-1)} \tag{5}$$

By extending (1) with the parameter b and the weight for previous decisions $w_{j(i-1)}$ we receive (5), which can be used for all subsequent decisions. Equation (5) generates unique weights for every single item and for each rater, while their performance on previous items is taken into account. Thus, not everyone is seen as equal and expertise can significantly increase ones impact on a vote, but at the same time drastic underperformance on previous items is considered and can in turn be used to penalize experts that may not perform well on this specific task. In the same manner, a novice may receive high scores due to excellent performance, hence receiving more weight than an expert. This approach allows flexibility within each vote, without being overly biased towards any specific group of individuals. Ideally, this yields an improved performance, which can be indicated by an increased amount of cases where a specific label can be derived from a majority vote. Items that have been seen as ambiguous can now receive a label because a smaller group of individuals with higher overall performance and expertise may swing the vote in their favor.

V. Preliminary Experiments and Results

Currently, it is planned to label a corpus consisting of music related texts from three different languages with 1000 texts per language. The labeling task will be done by 90 raters with different language levels. As the development of this corpus is still in progress, in this paper, preliminary results using an already existing data set including 3000 texts from the same domain labeled by a total number of 48 raters, whereas each text was rated by 20 to 27 raters, are presented. This dataset includes besides the text, the individual rating of each rater and the time stamp referring to the time at which the text was presented to the rater. Using these time stamps for successive texts, the response time was estimated. However, this data set does not include all features proposed in Section III.

For each text there were four possible cases: In the first case, the majority of raters decided that the text does not deal with music, thus giving it the label “not Music”. In the second case, the majority choose the middle ground or an uncertain outcome, meaning the item may contain music elements yet it is not enough to strongly identify it as such. This case is referred to as “uncertain”. Analogous to the first case, it might be possible to identify a majority that voted for music content, giving a text the label “is Music”. Lastly, if aforementioned cases all fail due to a lack of significant agreement for the specific text, e.g., one third voted “is Music” one third voted “uncertain” and one third voted “not Music”, then we receive the label “no Majority”, leaving the text as ambiguous and hard to identify. Especially, the last case is of relevance as the number of “no Majority” occurrences determines the performance of a given parameter set.

A first baseline is received by conducting unweighted majority votes, utilizing the labels that have been provided by the data set. In order to prevent unrepresentative results, we repeated the calculation with different thresholds for the entire data set of 3.000 unique text items. The majority threshold is defined in the interval of $[0.5, 0.95]$ with steps of 0.05. This way, it is possible to determine the interval for which the best performance is achieved. Afterwards, majority votes utilizing the weighted approach were made using the same thresholds. A comparison of the results can be seen in Table I. It becomes evident, that there are two fundamental ways of reducing the “no Majority” count and thus increasing overall agreement. As expected, with a less restrictive majority threshold, the cases with no majority agreement are reduced. Furthermore, Table I shows that the weighted approach presented in this paper leads to a further decrease in “No Majority” cases if the threshold is kept stable.

TABLE I. COMPARISON OF MAJORITY OCCURRENCES, FOR DIFFERENT THRESHOLDS, WITH AN UNWEIGHTED (UW)AND WEIGHTED (W) APPROACH.

Threshold	is Music		Uncertain		not Music		No Majority	
	UW	W	UW	W	UW	W	UW	W
0.5	1534	1560	3	6	1344	1357	119	77
0.55	1456	1481	2	2	1298	1302	244	215
0.6	1373	1410	2	1	1240	1249	386	339
0.65	1278	1314	0	0	1175	1182	547	504
0.7	1128	1191	0	0	1115	1139	757	670
0.75	982	1072	0	0	1064	1081	954	847
0.8	825	918	0	0	998	1025	1177	1057
0.85	648	739	0	0	931	952	1421	1309
0.9 0	435	504	0	0	821	842	1744	1654
0.95	236	270	0	0	601	636	2163	2094

Figure 1 describes the discrepancy of “no Majority” cases for both, the original unweighted data and our weighted approach utilizing different thresholds. It becomes imminent that we receive fewer cases of “no Majority” with the new setup for all tested thresholds. Furthermore, by looking closely at the data we can see that our weighted approach performs especially well with majority thresholds in the interval of $[0.7, 0.85]$. Peak performance was acquired at a threshold of 0.8. Regardless of the seemingly ideal interval, we maintain a steady improvement of 10% or more for all tested thresholds. An increase beyond 10% is quite valuable, especially when taking into account that our data set did not include sufficient data for all the features that we implemented.

Since the data set did not include data for all required features, some aspects had to be artificially created. First of all, for the Intra-Rater-Agreement, the necessary values were drawn from a normal distribution. The same procedure was used to get the values for the responses for the questionnaire, which included questions regarding the topic expertise. Furthermore, for the conscientiousness scores a normal distribution was used as well, yet a mean of 0.675 was assumed instead of 0.5 because of the wide availability of “Big Five Test” evaluations such as the one by van der Linden [19]. Finally, if a rater takes a break during the labeling process we may see a spike in the response time that could skew our data set in an unfortunate way. Thus, we shortened all response times longer than 5 minutes to a maximum of 5 minutes. In this experiment, all feature weightings $\vec{\beta}$ were set to 1.

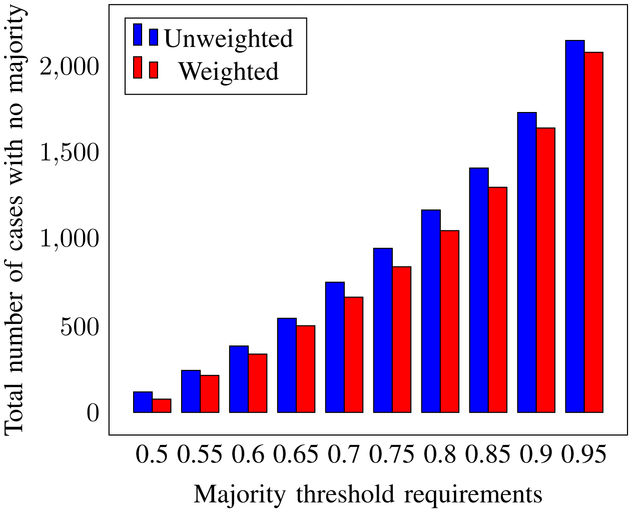

Figure 1. Comparison of “no Majority” cases for an unweighted and weighted majority vote with different majority thresholds.

Analogously to generating more majorities, this approach may also occasionally create disagreement. While the total “no Majority” count clearly decreases there could still exist individual cases in which the weighted approach derives the label “no Majority” whereas a classic unweighted majority vote may find a majority. It would be of great interest to see in which situations the weighted approach creates disagreement and if in some cases the majority in itself changes, meaning the unweighted majority labeled an item as, for example, “is Music” whereas the weighted majority labeled the same item as “not Music”.

VI. CONCLUSION AND FUTURE WORK

In this work, a flexible weighting approach for Inter-Rater-Agreement is proposed. Preliminary experiments have shown that using this approach may provide improved results for text labeling tasks even when missing data is artificially created. With actual available data the improvements might be even more significant since a correlation between response time and rater conscientiousness might exist.

We currently collect data to redo this experiment with non-generated feature data, which should provide more accurate insights. If the aforementioned or a similar correlation exists, the weighted majority vote may yield drastically improved results surpassing the current 10% mark. Additionally, in the next experiment we will evaluate the optimization potential by utilizing different settings for the weighting parameters β_i which were set to 1 in this first research phase.

Furthermore, similar experiments should be conducted using varying feature spaces to not only ensure the validity of this approach but also to discover potential performance variations between individual features. Finally, while in this study we only resort to rater dependent features, there is also the option to include item dependent features such as the item specific competence level of the rater.

REFERENCES

[1] D. Yvonne, A. Eva, and G. Gunnar, “Inter-Rater Agreement Using the Instrumental Activity Measure,” Scandinavian Journal of Occupational Therapy, vol. 7, 2000, pp. 33–38.

[2] R. Artstein and M. Poesio, “Inter-Coder Agreement for Computational Linguistics,” Computational Linguistics, vol. 34, no. 4, Dec. 2008, pp. 555—596.

[3] W. A. Scott, “Reliability of Content Analysis: The Case of Nominal Scale Coding,” Public Opinion Quarterly, vol. 19, 1955, pp. 321–325.

[4] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” Educational and Psychological Measurement, vol. 20, no. 1, 1960, pp. 37–46.

[5] J. L. Fleiss, “Measuring nominal scale agreement among many raters,” Psychological Bulletin, vol. 76, 1971, pp. 378–382.

[6] J. Cohen, “Weighted Kappa - Nominal Scale Agreement with Provision for Scaled Disagreement Or Partial Credit,” Psychological Bulletin, vol. 70, Nov 1968, pp. 213–220.

[7] K. Krippendorff, Content Analysis: An Introduction to Its Methodology. Sage, 2004, ISBN: 978-07-6191-545-4.

[8] E. A. Kolog, C. S. Montero, and E. Sutinen, “Annotation Agreement of Emotions in Text: The Influence of Counsellors’ Emotional State on their Emotion Perception,” 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016, pp. 357–359.

[9] R. Artstein, “Inter-Annotator Agreement,” Handbook of Linguistic Annotation, 2017, pp. 297–313.

[10] M. Boguslav and K. Cohen, “Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing,” Studies in Health Technology and Informatics, vol. 245, Jan 2017, pp. 298–302.

[11] R. Passonneau and B. Carpenter, “The Benefits of a Model of Annotation,” Transactions of the Association for Computational Linguistics, vol. 2, Dec 2014, pp. 311–326.

[12] A. P. Dawid and A. M. Skene, “Maximum Likelihood Estimation of Observed Error-Rates Using the EM Algorithm,” Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 1, 1979, pp. 20–28.

[13] O. P. John and R. W. Robins, “Accuracy and Bias in Self-Perception: Individual Differences in Self-Enhancement and the Role of Narcissism,” Journal of Personality and Social Psychology, vol. 66, Feb 1994, pp. 206–219.

[14] L. R. Goldberg, “The development of markers for the Big-Five factor structure.” Psychological Assessment, vol. 4, no. 1, 1992, pp. 26–42.

[15] M. R. Barrick and M. K. Mount, “The Big Five Personality Dimensions and Job Performance: A Meta-Analysis,” Personnel Psychology, vol. 44, no. 1, 1991, pp. 1–26.

[16] T. A. Judge, C. A. Higgins, C. J. Thoresen, and M. R. Barrick, “The Big Five Personality Traits, General Mental Ability, And Career Success across the Life Span,” Personnel Psychology, vol. 52, Dec 2006, pp. 621–652.

[17] L. Satow, “ B5T - Psychomeda Big-Five-Persönlichkeitstest,” Personnel Psychology, 2011.

[18] C. of Europe, “Common European Framework of Reference for Languages: Learning, teaching, assessment,” 2001.

[19] D. van der Linden, J. te Nijenhuis, and A. B. Bakker, “The General Factor of Personality: A Meta-Analysis of Big Five Intercorrelations and a Criterion-Related Validity Study,” Journal of Research in Personality, vol. 44, March 2010, pp. 315–327.