Reliability and Agreement on Category-Valence Classifications of Headlines

By Nico Nuñez, Andrew Duchon, and Phil Duong at Manzama, Inc.

Talking Points

- Insights classifies articles with a neural network model trained with human data.
- This model can be trained to agree with humans only as far as humans agree with themselves. Two humans will give the same category (a.k.a. subfactor) and valence for a headline only 87% of the time.
- The Insights models are within 10% of this maximum level of agreement.
- We now show only the factor and valence (not subfactor) which is correct 90% of the time.
- Every week, new topics show up in the news that the model has never been trained on, in which case it makes its best guess.
- We are constantly providing the model with new training data to tell it how to classify these new news articles.

Executive Summary

The classifications of articles provided by Manzama Insights are produced by a neural network model trained with human data. The classification relates the text of a headline to a category-valence combination. Because new events and terms are constantly emerging, we have an ever evolving 35-page guideline to help the human evaluators make consistent classifications. Nevertheless, given the rich complexity and variability of language, these classifications can be extremely difficult even for well-educated humans. Furthermore, "agreement" between human evaluators acts as the functional limit and aspirational goal for the model.

For instance, the most any two human evaluators agree with each other is 92% for Insights' 7 Factors, its most general level of classification. However, as the complexity increases, so does the disagreement rate between humans. At 31 Categories or Subfactors, the agreement drops to 89%, and drops again to 83% when looking at all 81 Category-Valence combinations. To account for the disagreement rates, the Gold Standard assignment was created for each headline possibility. Relative to the Gold Standard, the model reaches 86% agreement for Factors, 82% for Categories, and 74% for the full Category-Valence combinations. This indicates that the model is within 7, 9, and 12% of the human agreement rate, respectively.

Because some differences in the assignment are worse than others, an analysis of relative agreement vs. absolute agreement was required. Assessing the models' performance using the Krippendorff Alpha agreement metric, we see the upper bounds which humans relatively agree with each other on the Categories is 89%, and the most complex Category-Valence combinations is 87%. When the models are compared to the gold standard, we see 83% and 78%, respectively[1].

In conclusion, we see an 87% relative agreement and 83% absolute agreement rate between humans at the catval level, representing the upper limit and aspirational "accuracy" rate for the catval models. Analyzing the models' output, we see a 78% relative agreement and 74% absolute agreement rate with respect to human evaluators, so they are within 10% (78/87) of the maximum performance possible. In practical terms, one can expect the classification of about 1 in 4 articles to seem to be "off" and about 1 in 5 articles to be jarring. But that is at the catval level. Therefore we are instituting changes to the UI to show only factors, and perhaps only valence for the model's most confident predictions. In this manner, only 1 in 10 articles will appear to be misclassified.

That said, new topics are always in the news and until human data is provided on those topics, we don't know what the maximum level of agreement is. Therefore, we are constantly providing the model with new data to deal with these new topics.

[1] The factor level is omitted from the relative agreement discussion because the relative agreement metric is the same as the absolute agreement metric for factors since any factor is equally 'distant' or 'different' from all other factors. However, at the category and catval level this is not the case, since two categories or catvals can be part of the same factor, and thereby be less distant than two categories or catvals that are part of other factors.

Study

The reliability of the teaching signal in a machine learning model is fundamental for its ability to learn. In Manzama Insights' case, the model (or classifier) is trained to learn the relationship between article headlines (input) and "catvals" (output), which are Category-Valence combinations. The teaching signal is provided by headline-catval pairs (input-output pairs) that human evaluators have prepared for the model's training. Because of the inherent complexity and variability of language, the task of determining the correct catval for a specific headline can be hard, even for human evaluators. This opens up a few very important questions: how often do human evaluators agree on the chosen catval? How often do humans agree with the classifier catval prediction? The rate at which humans agree on a catval for a given headline will define the upper-bound of the "accuracy" for Manzama's machine learning model.

Even though the catval is the teaching signal used to train the neural network, there are three levels of granularity for each headline evaluation. The most complex and granular level is the aforementioned 'catval', comprised of one Category-Valence combination out of a set of 81 possible catvals. The second level of granularity is the Category, which is chosen from a set of 31 possible Categories. The most general level is the Factor, which is chosen from a set of 7 possible Factors. For example, consider the headline "Court of Appeal Rules Against Apple in Data Breach Case": the catval would be "Cyber Issues - Negative', the category would be "Cyber Issues", the valence would be negative as it relates to the subject-matter company (target company=Apple), and the factor would be "Operations". An outline of the possible evaluations and their different levels of granularity is presented in Appendix A.

In order to teach the machine learning model relationships between the words presented in a headline and the corresponding catval classification, each piece of training data should ideally be consistent and in agreement with all other pieces of training data. In order to increase consistency and diminish disagreement between training examples, Manzama developed a system called the Gold Standard, in which a specialized human reviews sets of human-generated training examples that are in disagreement, and determines the "best" catval for that controversial headline. The Gold Standard thereby resolves disagreements in evaluations and consequently trains the model with more consistent training data.

This study is focused on assessing interrater reliability: a measure of the extent to which raters, or human evaluators, agree on a training classification. There are various forms of interrater reliability metrics, but in this study we will investigate two: percent agreement and Krippendorff's Alpha.

Percent Agreement

Percent agreement is defined for any two evaluators as the rate at which these two evaluators agree on a headline-catval pair. Upon doing this analysis, results suggest that catval percent agreement between human evaluators ranges between 73% and 83%, depending on the pairs of human evaluators that are being analyzed[2]. For instance, on catval classifications, User 1[3] and User 3[4] agreed 73% of the time, while User 1 and the Gold Standard agreed 83%[5] of the time. When observing human evaluators' agreement on Category only, excluding valence, the range increases to 78%-89% agreement. When observing human evaluator's agreement on Factor, the broadest classification level, the range increases to 85%-92% agreement.

The highest level of human agreement attainable at the catval level (83%) is about 12% higher than the level that humans agree with Insights Classifiers overall. For instance, User 1 and Insights Classifiers agree on 76% of catval classifications, User 3 and Insights Classifiers agreed on 68% of them, and the Gold Standard and Insights Classifiers agreed on 74% of catvals.[6] In other words, the range of human-models agreement on catval classifications is 68%-76%. When taking into account all Classifier evaluations in our system, the rate of catval classification agreement with humans could at most increase by 9 percentage points, category classifications could increase 7 percentage points and factors by 6 percentage points.

---

[2] Some pairs of human evaluators have less than 73% agreement, but have smaller sample sizes

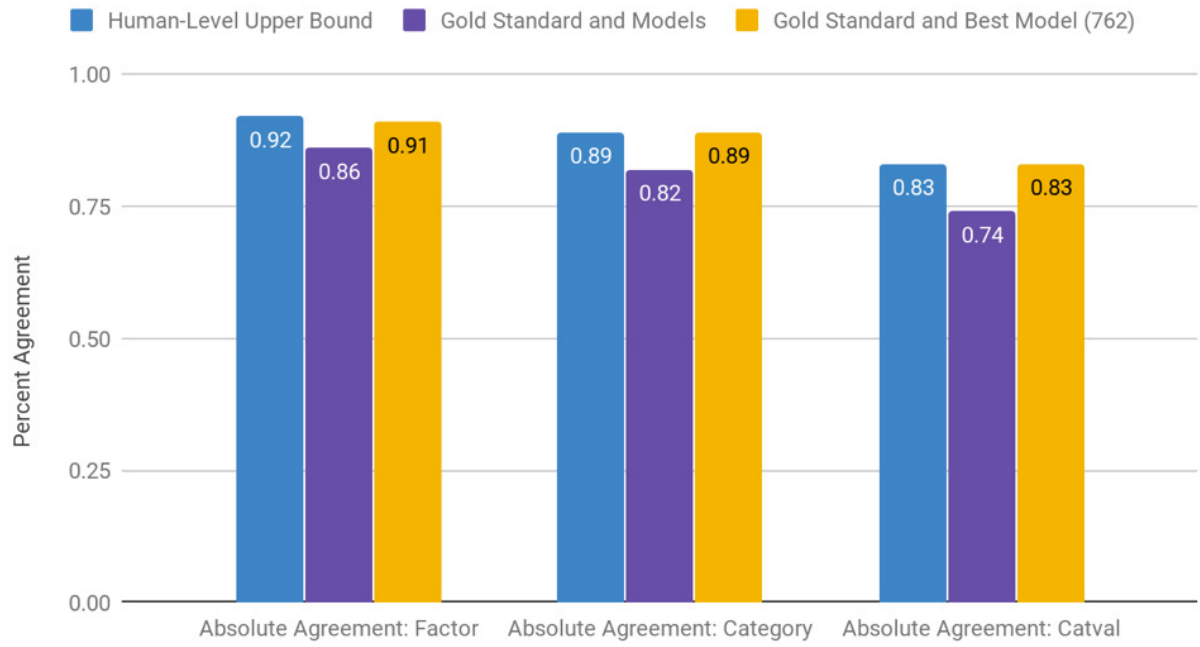
[3] Andrew Duchon, Co-Inventor, Director of Data Science

[4] Phil Duong, Director of Product Development

[5] User 1 is also the primary curator of the Gold Standard.

[6] A significant percentage of User 3's evaluations is in direct response to mis-classifications due of never-before-seen headline, event, or language combinations.

While this is true for Classifiers overall, one of our best models, Classifier 762, attains percentage agreement statistics that closely resemble the upper bound for humans. Classifier 762 agrees with the Gold Standard in 83% of catval classifications, 89% of category classifications, and 91% of factor classifications. Thus, Classifier 762 is within 1% of the human-level upper bound. While this is a significant accomplishment, new models must be deployed every month to keep up with the changes in the stories presented in the news. Therefore, the purple bar presented above represents a more accurate and reasonable estimate of expected percent agreement for a future model.



Percent agreement is an intuitive and easily-interpretable metric to quantify reliability. However, it has three considerable limitations. The biggest limitation is that it does not take into account chance agreement, that is, the rate of agreement that would ensue if evaluators were simply guessing their evaluations randomly (McHugh). The second biggest limitation is that only pairs of evaluators can be compared, and there are several evaluators providing training signals for the model at Manzama. The third biggest limitation is that percent agreement is binary: evaluators must either agree or disagree on the catval: there is no intermediate option.

Krippendorff's Alpha

Enter Krippendorff's Alpha: this interrater reliability statistic (1) computes an agreement score for any number of evaluators, (2) is not binary in the sense that it allows for the inclusion of a difference function across catvals, such that we could penalize less if evaluators agree on category but not valence, and (3) Krippendorff's alpha considers expected agreement, thereby taking into account chance agreement. Like the percent agreement statistic, the value of Krippendorff's alpha is contained between [0 ± sampling errors - systematic disagreement] and 1 (Krippendorff 2011, 1). $\alpha <$ 0 means systematic disagreement, $\alpha = 0$ means complete chance agreement and $\alpha = 1$ means complete agreement. For a more detailed explanation of the intuition behind this metric, refer to Appendix B. The formula for Krippendorff's alpha is presented below.

$$\alpha_{Krippendorff} = 1 - \frac{Observed\ Average\ Disagreement}{Expected\ Average\ Disagreement} = \frac{Observed\ Average\ Agreement - Expected\ Average\ Agreement}{Expected\ Average\ Disagreement}$$

The difference function takes as input two classifications and outputs a number between 0 and 1 that specifies how different those classifications are. We created a difference function that particularly penalizes valence disagreement: claiming valence is Positive when someone else believes it is Negative is considered 5 times worse than claiming valence is Neutral when someone else believes it is Negative or Positive. The specification of the difference function is presented below in Table 1.

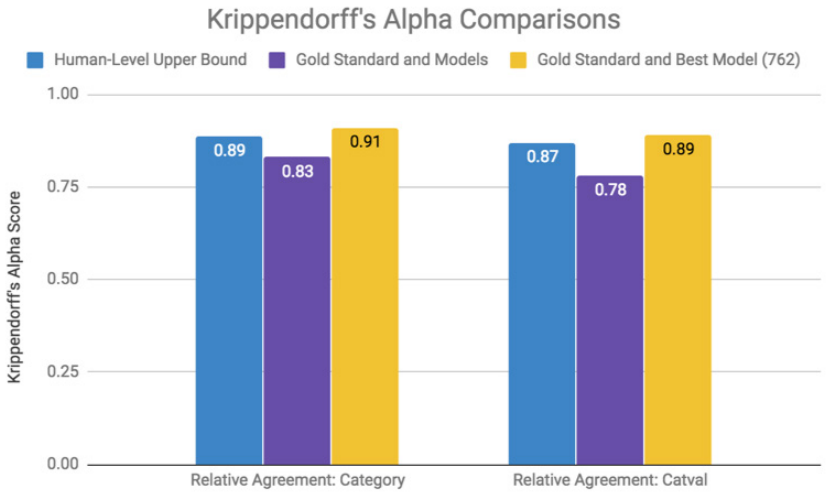
Factor	Category	Valence (Domain: POS, NEU, NEG)	Difference Function Output
Agree	Agree	Agree: Equal	0
Agree	Agree	Disagree: Opposite (i.e. POS vs. NEG)	1
Agree	Agree	Disagree: Not opposite (i.e. POS or NEG vs. NEU)	0.2
Agree	Disagree	Agree	0.8
Agree	Disagree	Disagree (Opposite or Not Opposite)	1
Disagree	Agree/Disagree	Agree/Disagree	1

Table 1. Difference Function for any pair of catvals

Krippendorff determines that a score of $\alpha = 0.800$ draws the line for supporting scholarly arguments, and $\alpha = 0.667$ is still acceptable (Krippendorff 2004, 12).

The Krippendorff Alpha relative agreement score at the catval level for Insights Classifiers is 11% below the maximum relative agreement score for humans, while the score at the category level is 7% below the maximum score for humans. Even though the relative agreement score for the Models is lower than the upper bound for humans, the difference is not as large as the difference in the absolute agreement scores. This is because there is more partial agreement within the 26% absolute disagreement between the Models and the Gold Standard than partial agreement within the 17% absolute disagreement between humans. In other words, there are many disagreements between the Models and the Gold Standard that are actually very close, just barely different. There are not as many disagreements that are very close between humans.

When observing the Gold Standard's agreement levels with one of our best classifiers, shown in yellow in the chart, we can see that the Krippendorff Alpha agreement score surpasses the upper bound for humans by 2%. This suggests that amongst the disagreements between Classifier 762 and the Gold Standard, many are correct at the factor level but disagree at the category level. Human-level disagreements, on the other hand (in blue), are different at the factor level at a higher rate than the disagreements between Classifier 762 and the Gold Standard. Hence, Krippendorff's Alpha metric allows us to distinguish between full-on disagreement, which implies a different factor, category and/or valence, from relative disagreement, which can imply the same factor, but different category or valence. However, as mentioned earlier, it would not be prudent to expect future models to obtain a Krippendorff Alpha score that matches one of our best models. Instead, a better estimate for the Krippendorff Alpha relative agreement score of a future model with the Gold Standard is presented in purple below.



In conclusion, we see it is relevant to distinguish between absolute agreement and relative agreement. Absolute agreement is easily interpretable: humans are in complete agreement up to 12% more often than the Gold Standard and all Insights Classifiers. Relative agreement allows us to distinguish between perfect agreement and partial agreement: humans reach partial agreement upto 11% more often than the Gold Standard and all Insights Classifiers. However, upon analysis of the best Insights Classifiers as of February 2019, we observe that the rates of agreement reached with the Gold Standard, both absolute and relative, match the highest recorded rates of agreements between humans.

Future Work

In future research, we hope to look into rates of different configurations of disagreements, instead of aggregating them into one Krippendorff Alpha metric. We aspire to obtain a set of 9 disagreement rates:

1. Factor agree, category agree, valence opposite
2. Factor agree, category agree, valence neutral
3. Factor agree, category agree, valence same (perfect agreement)
4. Factor agree, category disagree, valence opposite
5. Factor agree, category disagree, valence neutral
6. Factor agree, category disagree, valence same
7. Factor disagree, category disagree, valence opposite (perfect disagreement)
8. Factor disagree, category disagree, valence neutral
9. Factor disagree, category disagree, valence same

The most significant disagreement types are, as numbered above: 1, 4, 7.

In addition, we look forward to analyzing the absolute and relative agreement metrics for specific factors, categories and catvals. For instance, we would develop a ranking of the most contentious categories. This could be useful to pinpoint which sections of the Guidelines require more detail or which categories need to be explained again to specific evaluators in order to increase the agreement across humans.

Appendix A: Classification Granularity Levels

Factor: General Level	Financials	Government	Partners & Competitors	Operations	Products & Services	Management	Ignored
Category	Analyst	Politics	Competition	Attacks & Disasters	Intellectual Property	Executive Movement	Conference
Category	Bankruptcy	Regulation	Deals	Cyber Issues	Product	Executives	Crime
Category	Financials	Taxes	Mergers & Acquisitions	Expansion & Contraction	Product Liability	Insider Transactions	Marketing
Category	Stock News			Labor	Public Sentiment	Misconduct	Non-English
Category				Supply Chain	Sales	Share- holders	Non-Target
Category							Spam

There are 3 headline classification groupings. The broadest, most general grouping is the factor, shown in blue in the above table. There are 7 factors in total. Each factor can be broken into a set of categories, presented in green and purple below the factors in the table above.

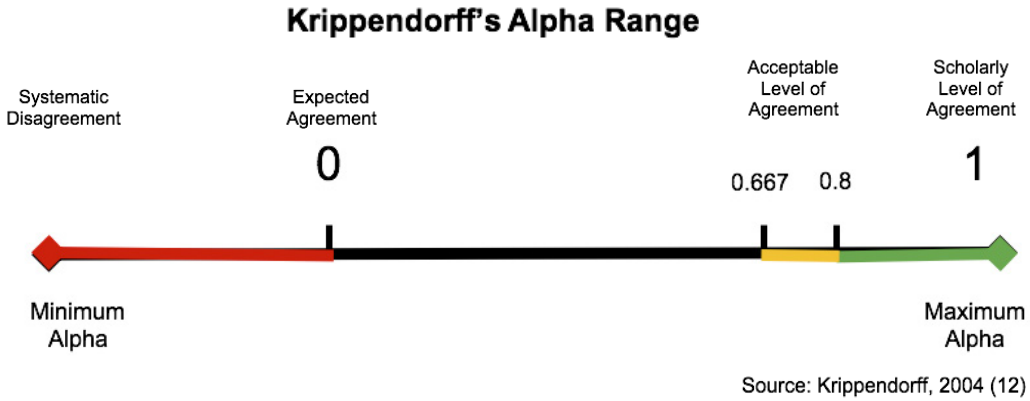
Some factors have more categories than others: the Ignored factor, for example, is broken into 6 categories, while the Government factor is broken up into only 3 categories. There are a total of 31 categories: this is the second most general headline classification grouping.

Appendix B: Krippendorff's Alpha Intuition

The Krippendorff Alpha metric is mathematically very different from percent agreement. It is a ratio of two values: the 'above-expected' percent agreement, and the expected percent disagreement. The above-expected percent agreement is the fraction of all pairs of evaluations that are in agreement (that are the same) beyond those that we would expect by random chance.

$$\alpha Krippendorff = \frac{Above\ Expected\ Percent\ Agreement}{Expected\ Percent\ Disagreement}$$

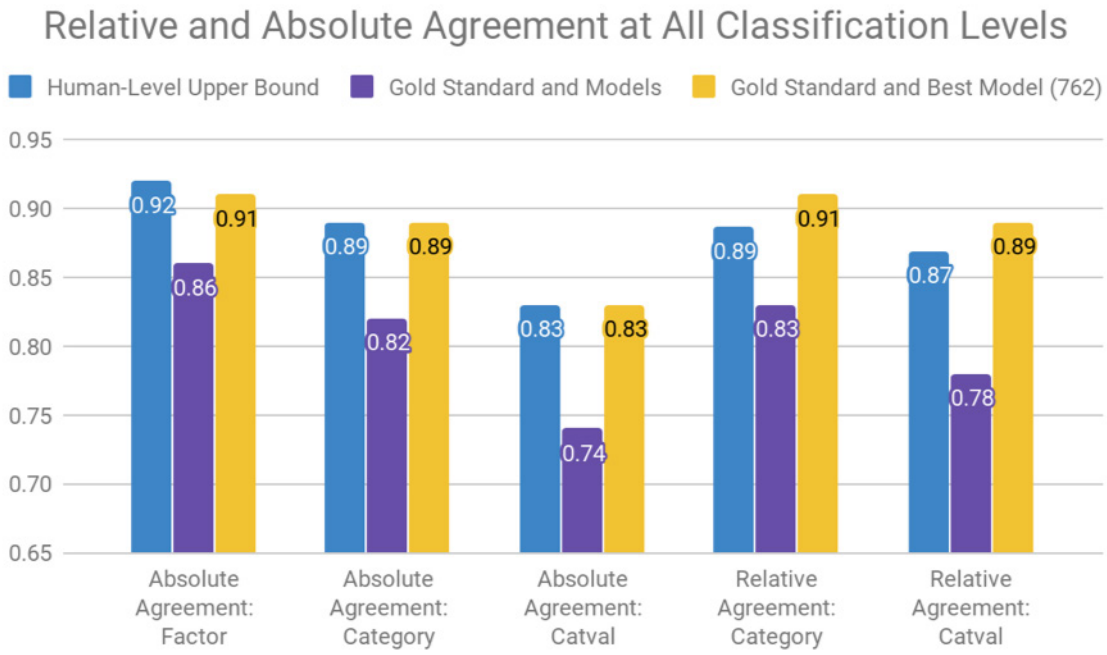
It is important to understand the difference between observed agreement and expected agreement: observed agreement is the percentage of all pairs of evaluations that were found to be the same, while expected agreement is the percentage of all pairs of evaluations that we expect would be the same if all humans were making evaluations by guessing randomly.



Source: Krippendorff, 2004 (12)

The expected percent disagreement is always larger than the expected percent agreement, because it is much more likely to disagree than it is to agree by randomly guessing. For this reason, the maximum Krippendorff alpha value is 1, and it occurs only when there is 100% observed agreement. When there is 0% observed agreement, the Krippendorff alpha value is at its minimum, which is a negative number. When observed agreement matches expected agreement, Krippendorff's Alpha obtains a value of 0.

Appendix C: Data



Sources

Data. Retrieved February 12th, 2019. https://docs.google.com/spreadsheets/d/1rt4IjmMK_q3XJ9wtoAVU_5qykVkJ3x11Z1WGbF8cF9E/edit#gid=229517561 Organized Data tab.

McHugh, Mary L. (2012) Interrater reliability: the kappa statistic. Biochemia medica vol. 22,3 : 276-82. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from http://repository.upenn.edu/asc_papers/43

Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. Human Communication Research, 30 (3), 411-433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x