Automation-coerced, increased dilution in correlation

Chung-hong Chan[1]

[1] Mannheimer Zentrum für Europäische Sozialforschung, Universität Mannheim, Germany

Abstract

Automated data-making methods in content analysis —like all measurements— are fallible. The purpose of this simulation study is to show this fallibility can lead to the correlation dilution effect: the biased estimation of true effect size towards zero, or, in other words, the unexpected reduction in statistical power. An alternative way to measure the performance of automated procedures, which focuses on the retention of statistical power, is proposed. This paper ends with best practices regarding planning, executing, and reporting of automated content analyses.

Keywords: Automated Content Analysis, Correlation Dilution, Measurement Error, Statistical Power

Word count: 6847

Automation-coerced, increased dilution in correlation

The leaderboard paradigm

Automated data-making procedures such as dictionary-based methods and supervised machine learning are introduced to content analysis to replace the labor-intensive manual coding process. In methodological studies, these procedures are an end in itself and many of these procedures have been shown to approximate human coding, the currently available gold standard.

When mainstream communication scientists talk about a method "outperforms" another (e.g. Van Atteveldt, Velden, & Boukes, 2021; Barberá, Boydstun, Linn, McMahon, & Nagler, 2020), the yardstick is usually the same as the one used by computer scientists in machine learning benchmarks (Raji, Bender, Paullada, Denton, & Hanna, 2021) such as SemEval or ImageNet Challenge: how more accurate does the method approximates human judgment than other methods.[1] In this so-called leaderboard paradigm (Ethayarajh & Jurafsky, 2020), the performance of a method is captured by *performance metrics* such as F1, accuracy, precision, and recall. Previous studies that validate automated procedures for content analysis reported and ranked metrics such as F1 from various methods (e.g. Van Atteveldt et al., 2021; Barberá et al., 2020; Dobbrick et al., 2021). The process of validation, as captured in Grimmer and Stewart (2013)'s "validate, validate, validate" motto, is mostly about the whether the output of a procedure approximates human understanding of meanings in the source materials, otherwise known as semantic validation (DiMaggio, Nag, & Blei, 2013).

Validation and revalidation of automated procedures are of paramount importance and many studies have so far demonstrated the domain specificity of many of these

[1] A minority of communication scientists argues also for the importance of alternative yardsticks, such as cost (Guo et al., 2019), interpretability (Dobbrick, Jakob, Chan, & Wessler, 2021) and reproducibility (Chan et al., 2020).

procedures (Dobbrick et al., 2021; González-Bailón & Paltoglou, 2015). Van Atteveldt and Peng (2018) specify that "even if a researcher uses an existing off-the-shelf tool with published validity results it is vital to show how well it *performs* in a specific domain and on a specific task." (p. 87, emphasis added) I agree with Van Atteveldt and Peng (2018)'s conviction but propose not to look at the *performance* in the leaderboard paradigm. An alternative paradigm is proposed in the next section.

The power retention paradigm

I propose to look at the performance of automated procedures in an alternative paradigm called "(statistical) power retention" paradigm. This paradigm doesn't deal with how close the automated procedures approximate human coding *directly*. Instead, the power retaining paradigm deals with how effective automated procedures facilitate hypothesis testing by keeping the expected statistical power.

In a problem-driven content analysis (Krippendorff, 2018),[2] a data-making procedure such as manual coding or supervised machine learning is not an end in itself, but a means to meet an end. Quantitative Content analysis is always deductive, i.e. the ultimate goal of a content analysis is the theory-testing deductive procedure based on statistical inference (e.g. hypothesis testings) *after* the coding procedure (Krippendorff, 2018). Therefore, the actual performance of an automated procedure for content analysis should *not* be measured by how close the automated procedure approximates human coding, but by how useful the

[2] Please note that this article deals exclusively with (automated) content analysis with a problem-driven design, which by definition always aims to deductively answer some research questions (Krippendorff, 2018). There are also other designs: text-driven design (also known as qualitative content analysis) and method-driven design for justifying methods. Outside the realm of content analysis, less stringently defined approaches such as "text-mining", "text-as-data" approach, "QTA" (quantitative text analysis), "NLP" (natural language processing) or even marketing terms such as Big Data or AI analyses might not be deductive. "Text-mining", for instance, implies "mining" (or uncovering) patterns in data and thus should be entirely inductive.

automated procedure is to faciliate the downstream tasks such as hypothesis testing.[3]

Correlation dilution in content analysis

How can data-making procedures impact the subsequent hypothesis testing in a content analysis? To answer this question, one has to first understand what threats the validity of hypothesis testing. Sampling error and measurement errors are two of the most common threats. Both random and systematic sampling error has received much more attention by content analysts (e.g. the calculation of sample size beforehand and ensure the representativeness of the sample by random sampling, see the best practice article by Lacy, Watson, Riffe, & Lovejoy, 2015) and thus are not repeated here. Measurement errors, on the other hand, does not receive a lot of attention (except e.g. Bachl & Scharkow, 2017; Krippendorff, 2011). Even the reporting of reliability is now a standard procedure, the relationship to hypothesis testing is not explicitly explained (except Geiß, 2021, more on this later).

Applying the Classical test theory, the code of an article is an observed variable (O) and this observed result is a combined result of the unobserved true score (T), and measurement errors (E, i.e. $O = T + E$). The instrument used to obtain the observation O is *fallible* and such measurement errors E are unavoidable. And there are two components of measurement errors: random measurement error and systematic measurement error. Random measurement error, as the name implies, is the *unpredictable* variation and would only influence the variance of a variable but not the average level. The direction of influence is, however, predictable as random measurement error always increases the variance. In content analysis, interrater variability is a measurement error that is usually assumed to be random because coders are assumed to be interchangeable (Bachl &

[3] An analogy to this is the performance (or efficacy) of a new vaccine is measured by how useful it is to bring benefits to the patients, not by how close the new vaccine approximates the pharmacokinetic of the "gold standard", or any, existing vaccine.

Scharkow, 2017; Krippendorff, 2011).

On the contrary, systematic measurement error is the *predictable* variation that can be explained by systematic factors. One example of this systematic measurement error is mixing German and French articles together and training two different classification models for each individual language (Gilardi, Gessler, Kubli, & Müller, 2021). If the prediction models are different in performance, e.g. German articles are classified less accurately than French articles, the measurement error in this case is predictable by a systematic factor of whether or not the content is in German. Systematic measurement error could influence both the variance and the average level of a variable, but the direction of influence depends on the composition of data.

Let's make an unrealistic assumption that systematic measurement error is negligible. The direct consequence of doing statistical tests with observations containing random measurement error is the unexpected increase in variance. This inflated variance can lead to two related consequences: 1) the effect size estimation (e.g. correlation coefficient) is biased towards zero[4] and 2) when the sample size and α (Type I error rate) remain the same and the estimated effect size is biased towards zero, it unexpectedly increases β (Type II error rate). Figure 1 displays a simulation of this phenomenon. A correlated data of true X and Y are generated with a preset Pearson's correlation of 0.7. Then some random noise is added to X to simulate the observation of X with random measurement error. In the density curve, the observed X with random measurement error has the same peak as the true X but with a much broader spread, i.e. a higher variance. When compared with the true correlation, the correlation between the observed X and Y is biased towards zero, as indicated by the much flatter regression trend line.

---

[4] The most intuitive example is the calculation of Cohen's D, which is calculated as the standardized mean difference: $d = \frac{\bar{x}_1 - \bar{x}_2}{s}$. When the mean difference (numerator) remains constant, a larger variance increases the sample standard deviation s (denominator) and thus makes the effect size bias towards zero.

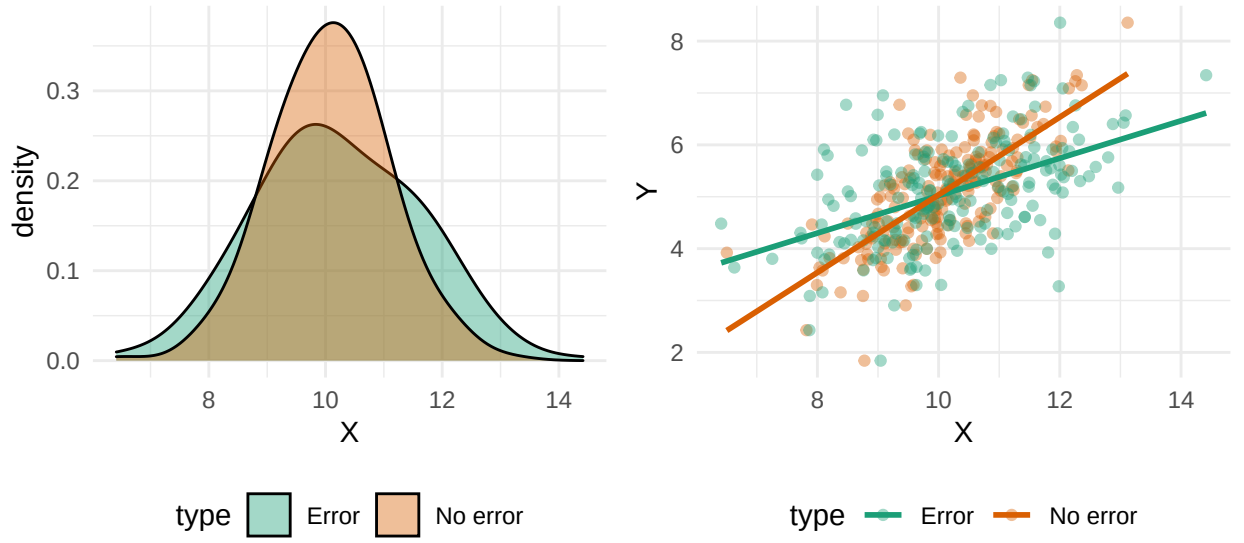

Figure 1. A simulation of correlation dilution: the inflated variance when observations containing random measurement error (Left) and the weaken correlation when observations containing random measurement error. (Right)

This phenomenon has long been discovered by statisticians (e.g. Spearman, 1904) and goes with different names, e.g. correlation dilution, correlation attenuation, attenuation bias, and measurement error bias. In this paper, I am going to call it correlation dilution.[5]

For content analysis, Geiß (2021) demonstrates the problem of correlation dilution with a series of Monte Carlo simulations, although the problem is not called correlation dilution in the paper. Coding accuracy (such as coder agreement), if one treats it as a random measurement error, can also induce the correlation dilution effect in content analysis. For example, the observed correlation ($\hat{r}$) in a content analysis, according to Geiß (2021)'s simulations, can be expressed using the following sigmoid function

[5] Please note that I have restricted the discussion to correlation but not regression; effect size but not regression coefficients. A related phenomenon called regression dilution is slightly more complicated. Similar to correlation dilution, the effect size of a regression model (R^2) on observations with measurement errors is also unexpectedly biased towards zero. When it's the independent variable with measurement errors, the slope of the regression model is also biased towards zero. But the slope of the regression model is not biased towards zero, if it's only the dependent variable with measurement errors (Fuller, 2009).

$$\hat{r} = \frac{1 \cdot |\rho|}{1 + e^{(-(b+d\cdot|\rho|))\cdot(Q-c))}} \tag{1}$$

where ρ is the true effect size, Q is the coding accuracy and b, c, and d are some constants. The denominator gets bigger when Q gets smaller, which results in the observed correlation getting weaker also. With this important finding, Geiß (2021) suggests that one should "consider coder agreement, sample size, and effect size in conjunction" (P. 86). Geiß (2021) situates his simulations in the realm of content analysis in general and automated content analysis is also a kind of content analysis. In the next section, I will talk about an issue that is unique to automated content analysis.

Automation-coerced, Increased Dilution in Correlation

Manual coding is an inevitable fallible procedure. Automated procedures such as dictionary-based methods or supervised machine learning are either validated with or trained on data generated by the fallible manual coding procedure. As manually coded data are considered to be the "gold standard" benchmark, automated procedures cannot have an accuracy rate larger than 100% and are assumed to be inherently "worse" than manual coding procedures. The disagreement between manual coding and automated procedures, which is also inevitable, creates another source of fallibility. The relationship between the reliability of manual coding procedure and the subsequent validation process of automated procedures has been demonstrated by Song et al. (2020). Concretely, the quality of human-coded data used for validating automated procedures are rarely checked and the quality can bias accuracy metrics such as F1.

The end products of automated procedures are what Knox, Lucas, and Cho (2022) called "(imperfect) learned proxies."[6] In the sentiment analysis scenario, for instance, the

[6] This is a direct quotation of the term, not an endorsement. It is important to note that the end products

surrogate of "sentiment" is learned from the textual content. Except in a handful of articles in which the actual sentiment are manually observed to serve as the so-called "gold standard", we never "observe" any sentiment. Instead, we infer the sentiment from the observed textual content using a certain prediction model. If having measurement error means imperfect in Knox et al. (2022)'s sense, then the end product of the automated procedure using in an automated content analysis should be called "imperfect imperfect surrogate". It is because the original "gold standard" is known to be imperfect too (Geiß, 2021; Song et al., 2020). An automated procedure based on a prediction model with 100% accuracy in predicting the imperfect human coding is only equivalent to the imperfect human coding. But in reality, the accuracy of the prediction model —the "performance" of a model in the leaderboard paradigm— can hardly be 100%. Therefore, automated procedures contain two sources of measurement error: 1) the inaccuracy of the prediction model to predict human understand and; 2) the inherent interrater disagreement. Two wrongs (measurement errors) certainly don't make one right. Instead, having two sources of measurement error inflates the variance further and dilutes the true correlation even more. This **a**utomation-*i*nduced, **c**oerced **di**lution *in* **c**orrelation (ACIDIC) is probably unique to automated content analysis. Knox et al. (2022) observe that it is a "common practice of conflating proxies with the underlying true concept". I echo this sentiment. I rereviewed the 37 communication research articles reviewed by Song et al. (2020) of having validated automated content analysis, none of the research papers acknowledges this ACIDIC problem.

---

of automated content analysis should not be called proxies because a proxy variable is defined as "a variable that is used in place of one that **cannot** be measured" (Upton & Cook, 2014, emphasis added). In the case of automated content analysis, one **can** certainly code the materials manually. Perhaps out of scalability concerns one doesn't want to code the materials manually. A more proper term for this should be "surrogate" which is defined as "a variable that can be measured (or is easy to measure) that is used in place of one that cannot be measured (or is difficult to measure)" (Upton & Cook, 2014).

Simulate ACIDIC in automated content analysis and measure performance holistically (pH)

Similar to the studies by Geiß (2021), Song et al. (2020), and Bachl and Scharkow (2017), this paper is going to study ACIDIC using Monte Carlo simulation. Building on top of Geiß (2021), the observed point and interval estimations of the observed correlation from an automated content analysis are governed by:

1. True effect size (ρ)
2. Sample size of the entire analysis (n)
3. Type I error rate (α)
4. Predictive accuracy (ζ)
5. Coder accuracy when coding gold standard for validation (Q)

Using only the items 1, 2, and 3, one can also calculate the expected statistical power ($1 - \beta$). Due to the ACIDIC induced by items 4 and 5, one would anticipate a reduction in statistical power. Therefore, the actual performance of an automated procedure in content analysis should be measured by how much expected statistical power it can retain. The power retaining performance is a holistic measurement of all 5 items in the above list; not just the item 4 as in the leaderboard paradigm. If an automated procedure cannot retain enough statistical power, it can't facilitate the deductive procedure and shouldn't be deployed for the purpose of automated content analysis. I will demonstrate how to measure the performance in the power retaining paradigm later in this article.

Methods

The purpose of the Monte Carlo simulation is to study how the all five items specified above (ρ, n, α, Q, and ζ) influence the observed effect size. There are many types of data one could study in a content analysis. In this simulation, the simplest case of 2×2 is used: there is a binary exposure variable X (e.g. being uncivil or not) and a binary outcome

variable Y (e.g. being shared or not). In this simulation, it is assumed that Y can be measured without any measurement error. X is only manually coded in a handful of the articles for training the prediction model, as well as evaluating the out-of-sample accuracy of the prediction. For the entire dataset, X is the output from the prediction model trained on some other observed variables (e.g. textual materials). Reversing the X and Y, i.e. the outcome variable is learned but not the exposure variable, will not change the result because this simulation only looks at observed effect sizes. However, when both X and Y are learned that would change the observed effect size drastically. However, it is uncommon for both exposure and outcome variables to be learned in an automated content analysis.

True effect size (ρ)

Similar to Geiß (2021), the correlation coefficient R is used. It is because most readers will be familiar with R. Although R presupposes interval- or ratio-level data, it is still possible to calculate an effect size that is equivalent to R from a 2×2 table. For this purpose, I calculated the mean square contingency coefficient, ϕ (Yule, 1912). This value is used as ρ and simulated data were generated with a prespecified level of ρ. In the following sections, ρ denotes the true effect size and R denotes the observed effect size for clarity.

The method to generate 2×2 data with a specific ρ is first selecting a random number from the uniform distribution $\mathcal{U}_{[0.1,0.5]}$ to denote the probability of the outcome, i.e. $P(Y = 1)$. Then, another random number from the uniform distribution $\mathcal{U}_{[0.01,0.5]}$ is selected to denote the probability of the outcome in the unexposed group, i.e. $P(Y = 1|X = 0)$. With the two probabilities $P(Y = 1)$ and $P(Y = 1|X = 0)$, the probability of the outcome in the exposed group, i.e. $P(Y = 1|X = 1)$, is found by the bisection method (Corliss, 1977) such that ϕ is equal to the prespecified level with less than 0.01 discrepancy.

With $P(Y = 1)$, $P(Y = 1|X = 0)$, $P(Y = 1|X = 1)$, and the total sample size,

simulated data are generated. Similar to Geiß (2021), simulated data with the following levels of ρ are generated: 0, .05, .1, .15, .2, .3, .4, .5, and .75. It is important to note that (observed) effect sizes found in communication research are usually in the lower end of this spectrum (Rains, Levine, & Weber, 2018). Effect sizes beyond .75 are extremely rare empirically.

Sample size of the entire analysis (n)

The following sample sizes are selected: 500, 1 000, 10 000, and 30 000. The upper limit is substantially larger than the one selected by Geiß (2021) (1 000) because automated content analysis usually has a much larger sample size. This range also covers most of the studies reviewed by Song et al. (2020). However, it is important to note that these are *sample sizes*, not *population sizes*. Automated content analytic studies with a data size larger than 30 000 usually study the entire population (e.g. Su et al., 2018 with a whopping data size of 243 235 637). In those cases, frequentist inference does not apply (Western & Jackman, 1994) even though p-values are still calculated in those population studies for no purpose.

Type II error rate (α)

The generally accepted level of 0.05 is selected.

Predictive accuracy (ζ)

In this simulation, ζ is assumed to be measured in the validation procedure. In the machine learning literature, it is confusingly called "test accuracy" (the accuracy in the test set). Here, I use the term "out-of-sample accuracy" from the statistical forecasting literature. The data used for validating the model are assumed to be coded similarly to the training samples and the data used for calculating Q.

There are many ways to report the predictive accuracy of a model. The most

commonly used metrics are Correct Classification Rate (CCR, also known simply as "accuracy") and F1. All of these measurements can be broken down into true positive rate (ζ_+), true negative rate (ζ_-) and prevalence of exposure ($P(X = 1)$). CCR can be expressed as:

$$CCR = P(X = 1) \times \zeta_+ + (1 - P(X = 1)) \times \zeta_- \tag{2}$$

F1 can be expressed as the harmonic mean of ζ_+ and positive predictive value (PPV, also known as "precision").

$$F1 = \frac{2 \times (\zeta_+ + PPV)}{\zeta_+ \times PPV} \tag{3}$$

Using the Bayes rule, PPV can be expressed as a function of ζ_+, ζ_-, and $P(X = 1)$.

$$PPV = \frac{\zeta_+ \times P(X = 1)}{\zeta_+ \times P(X = 1) + (1 - \zeta_-) \times (1 - P(X = 1))} \tag{4}$$

Similar to the case of ρ, random combinations of ζ_+ and ζ_- were generated with the fixed $P(X = 1)$ from the simulated data the specific level of CCR or F1. The following levels of F1 and CCR are selected: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.9.

Coder accuracy (Q)

The data used for calculated Q are assumed to be coded similarly to the training samples and the data used for calculating ζ.

As the impact of coder accuracy to the observed effect size has been well studied by Geiß (2021), the current simulation selects only one level, .7. Similar to the "Equality

Distribution" guessing mode, Q of .7 means 70% recognition and 30% guessing. When it is needed to guess a binary answer, it is based on a simulated coin flip. Therefore, for the 30% guessed answers, there are still 50% chance to be correct —what Krippendorff (2011) calls "agreement by chance".

Simulation

The entire simulation used nine levels of ρ, four levels of n, two different accuracy metrics, eight levels of accuracy, one level of Q and one level of α. In total, there are $9 \times 4 \times 2 \times 8 = 576$ scenarios. For each scenario, 1 000 simulation runs were made.

Each simulation run, which represents one empirical study, contains the following steps:

1. Generate n pairs of X and Y with $P(Y = 1)$, $P(Y = 1|X = 0)$, $P(Y = 1|X = 1)$ derived from the specific ρ
2. Manipulate X with ζ_+ and ζ_- derived from either the prespecified F1 or CCR. If $X = 1$, this X has $1 - \zeta_+$ chance to flip to 0. Similarly, if $X = 0$, this X has $1 - \zeta_-$ chance to flip to 1.
3. Manipulate X again with Q using the "Equality Distribution" guessing mode (Geiß, 2021)
4. Calculate R between the manipulated X and Y

For each scenario, 1000 observed effect sizes were calculated. The 50^{th} (median), 2.5^{th}, and 97.5^{th} percentiles of these observed effect sizes was calculated. The latter two correspond to the upper and lower limits of interval estimation when $\alpha = 0.05$.

Results

Figure 2 (F1) and Figure 3 (CCR) show the simulation results. In general, ACIDIC is evidenced in all cases. Take $\rho = .2$ as an example, the 97.5^{th} percentile of R never reaches

.2. Even with F1 = 0.9 (the accuracy level of some state-of-the-art machine learning model) and a large sample size, the upper limit of R is only around .1.

Similar to Geiß (2021), n does not change the trajectory of the relationship between median R and ζ when ρ is constant. n only changes the interval estimation. A higher n produces a narrower interval estimation, which guards against committing Type II errors.

The simulation also shows that ζ does not change the Type I error rate, a similar finding to Geiß (2021). It is evidenced by the flat line in all cases where $\rho = 0$, i.e. the null hypothesis is actually true. However, ζ does change the Type II error rate. When $\rho \neq 0$ (null hypothesis is not true), for instance $\rho = .2$, and $F1 \leq .80$, zero is still included in the interval estimation in all studied n. As a matter of fact, models with $F1 \leq 0.80$ cannot prevent Type II error, even when ρ is unrealistically large (0.75) and n is 30 000.

Measuring performance in the power retention paradigm

Suppose the expected true effect size ρ of a study is .0852 and n is 2019. In the validation study, the coder accuracy, as indicated by Krippendorf's α, is .8964. The ζ_+ and ζ_- of the trained prediction model are .8310 and .7210 respectively. Suppose the Type I error rate is 0.05.

In order to calculate the observed statistical power, a Monte Carlo simulation is performed and each simulation run contains the following steps:

1. Generate n pairs of X and Y with $P(Y = 1)$, $P(Y = 1|X = 0)$, $P(Y = 1|X = 1)$ derived from the specific ρ
2. Manipulate X with ζ_+ and ζ_- derived from either the prespecified F1 or CCR. If $X = 1$, this X has $1 - \zeta_+$ chance to flip to 0. Similarly, if $X = 0$, this X has $1 - \zeta_-$ chance to flip to 1.
3. Manipulate X again with Q using the "Equality Distribution" guessing mode (Geiß,

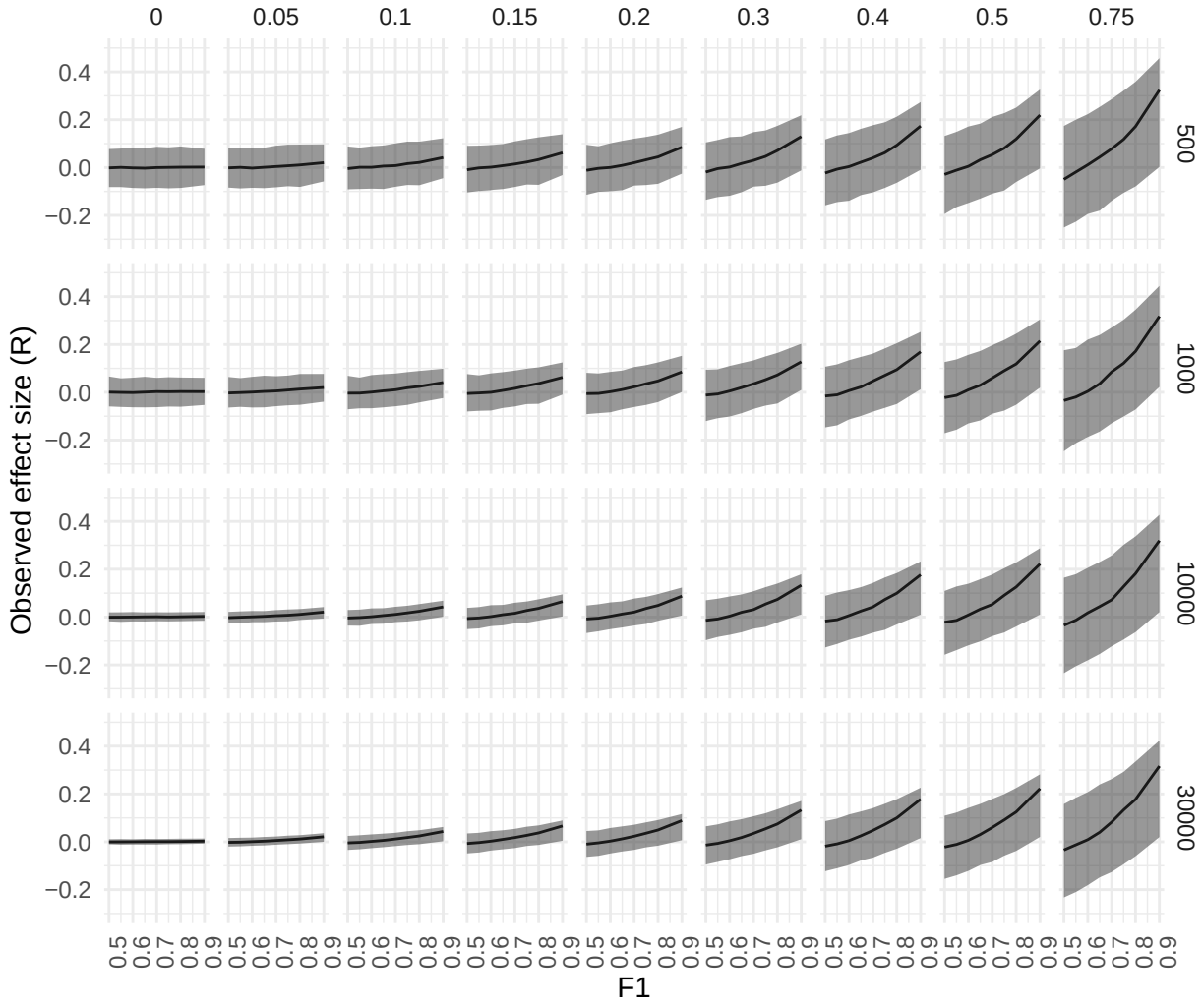

Figure 2. The relationship between F1 (X-axis) and observed effect size (Y-axis) by true effect size (column) and sample size (row). The line represents the median of all simulated observed effect sizes and the ribbon represents its 95% interval estimation.

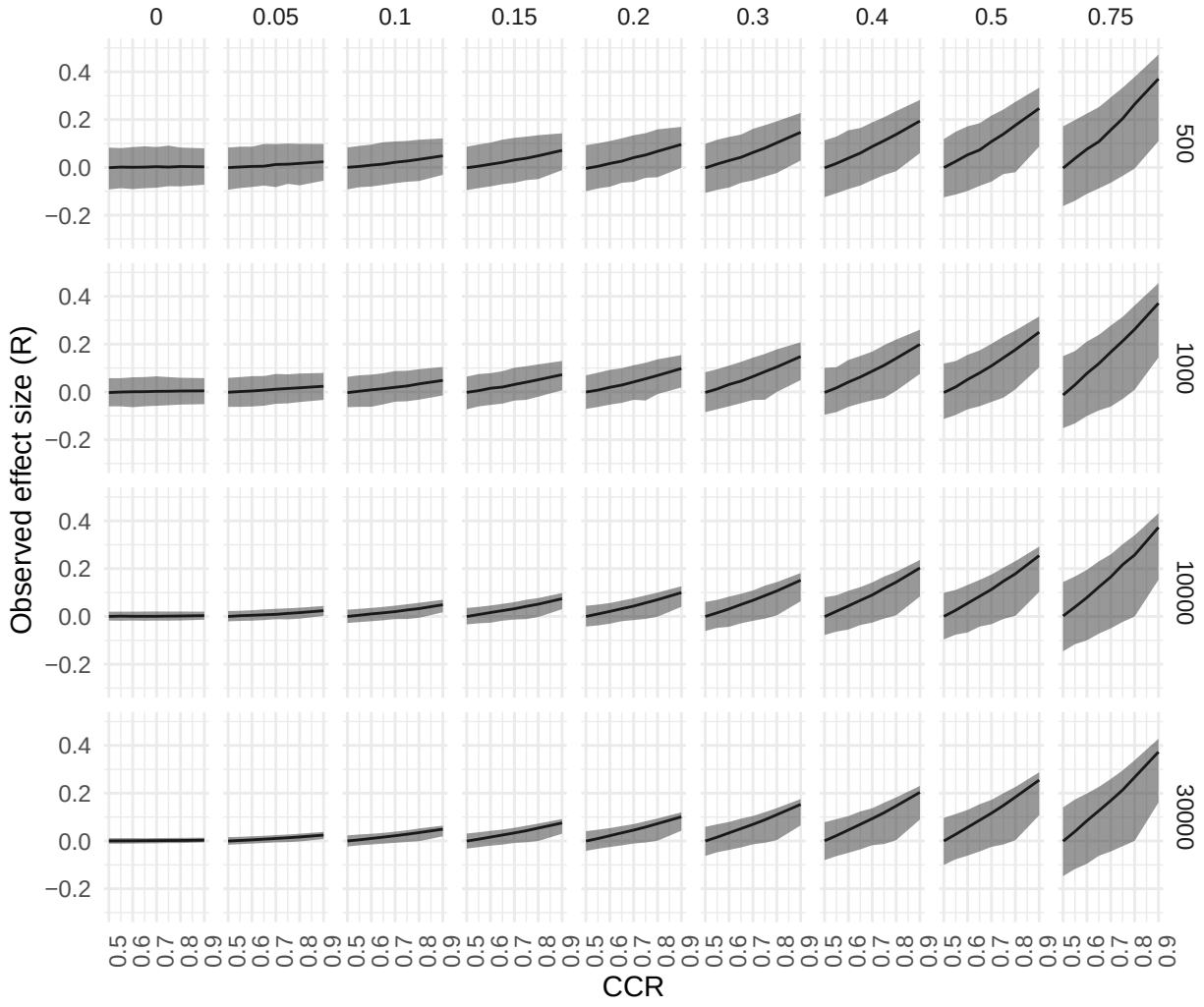

Figure 3. The relationship between correct classification rate (X-axis) and observed effect size (Y-axis) by true effect size (column) and sample size (row). The line represents the median of all simulated observed effect sizes and the ribbon represents its 95% interval estimation.

2021)

4. Calculate R between the manipulated X and Y
5. Calculate the (two-sided) P-value of R

β is equal to the total number of runs with $P > \alpha$ divided by the total number of runs. And the observed power $(1 - \beta)$ in this case is 41.1%.

Then, the expected statistical power is calculate. It is calculated by substituting ζ_+, ζ_-, and Q with 1. For most of the cases in automated content analysis, one can expect this power to be 100% and this step is not needed in practice.[7]

By comparing the observed and expected statistical powers, the statistical power reduces to 41.1% only due to ACIDIC. That is also the statistical power retained by the automated procedure. A lower power retention indicates less fit for use in content analysis. This retained power, dubbed Φ, can be calculated using the R function `cal_bigphi` provided (https://osf.io/vhwka/?view_only=84fda38794294dee8c34a2e563024045).

```
cal_bigphi(rho = 0.0852, n = 2019,
           tpr = 0.831, tnr = 0.721,
           Q = 0.8964)
```

Conclusion

The simulation study in this paper demonstrates the correlation dilution induced by the random measurement errors associated with automated procedures. This ACIDIC problem can drag down the planned statistical power. An R function is provided to calculate the power retained by automated procedures. The retained power Φ is a holistic

[7] For $\rho = .2019$, $n > 106$ would have 100% power. For a corpus of $n \leq 106$, it would be better off using manual coding, which only suffers from coding errors. If it is really needed, this can also be derived analytically using the central limit theorem because the distribution of R approximates normal when R is measured without measurement errors, i.e. $\mathcal{N}(\rho, \sqrt{\frac{1-\rho^2}{n}})$.

measurement of many factors: ρ, n, α, ζ, and Q.[8] For the purpose of automated content analysis, Φ is more meaningful than metrics such as F1 and CCR, which are used in the so-called leaderboard paradigm (Ethayarajh & Jurafsky, 2020) which focus only on ζ and are often reported in communication research (Song et al., 2020).

Having said so, Φ is not longer meaningful when n is getting too high. Any combination with $\rho \neq 0$, $\alpha > 0$, $\zeta_+ > 0.5$, $\zeta_- > 0.5$, and $Q > 0$ produces a 100% Φ, when n approaches positive infinity. It also appears that a straightforward solution to maximize the retained statistical power is to increase n, rather than improve either ζ or Q. First of all, this practice is not to "retain" more statistical power, but to **ob**tain. A procedure can only **re**tain more power than another procedure when comparing the two procedures with the same n, α, and ρ.

There is no shortage of digital data for us to analyze (remember the whopping data size of Su et al., 2018). These n articles do not need to be manually coded and therefore there is virtually no cost to analyze more digital data to obtain more statistical power. Is it fine to obtain almost infinite statistical power by boosting n?

As found in the current Monte Carlo simulation (as well as predicted by the law of large numbers), a higher n certainly reduces the chance of Type II error when ζ, Q, ρ, and α are constant. However, the point estimate of the effect size, as indicated by the median estimates in Figure 2 and Figure 3, do not change with n. An extremely bad predictive model with low ζ and Q still dilutes the correlation. When the null hypothesis is not true,

---

[8] I refrain from suggesting a cut-off point for Φ, not least because all cut-off points are controversial. Suppose you are a security analyst and world peace is at stake. You need to find, via automated content analysis, the peace-making signals in secret government cables from a geopolitical adversary. A mutual nuclear destruction will be triggered if a Type II error is committed. Would you accept a nice looking $\Phi = 0.80$ for this task? Many books teach us to calculate sample size with 80% power, right? Remember, there is a 20% chance of mutual nuclear destruction.

astronomical sample sizes can reduce the chance of obtaining a zero effect size. But they can't change the diluted correlation and make it even more certain to obtain a diluted correlation. **If the goal is to obtain an unbiased estimation of the true effect size, the only way is to improve both ζ and Q.** For population studies such as Su et al. (2018), this point is even more important because P-values therefrom are meaningless. The primary goal should be to estimate unbiased effect sizes. For those studies, it is more imperative to improve both ζ and Q.

Regarding the diluted correlation, corrective actions to adjust for it have been available since 1904 and continuously improved ever since (Frost & Thompson, 2000; Spearman, 1904). However, these methods are controversial as there is an overadjustment risk (Osborne, 2002). If these adjustment methods are to be used, they should be used exclusively for exploratory or sensitivity analyses (Hutcheon, Chiolero, & Hanley, 2010). In my opinion, the deployment of these adjustment methods should also be preregistered.

Instead of relying on these corrective actions, I advocate maintaining proper conduct. In an automated content analysis with proper conduct, the measurement errors are minimized and clearly documented. There are some excellent papers on how should (automated) content analyses be planned (Geiß, 2021), executed (Barberá et al., 2020; Van Atteveldt et al., 2021), and validated (Song et al., 2020). In light of the current finding, I would like to reinforce certain points from these papers by suggesting a few best practices.

1. Design automated content analysis carefully

Any deductive study needs a careful design, (automated) content analysis is no exception. Krippendorff (2018) (specifically, Chapter 14), Lacy et al. (2015), and Geiß (2021) introduce how content analyses should be designed. I would like to add just two points. First, automated content analysis is about making valid inferences. Selecting a valid mode of statistical inference should also be a part of the research design. Population

studies, which involve the entire population of text, do not allow frequentist inference. Alternative paradigms are needed: descriptive (Gerring, 2012), Bayesian(Western & Jackman, 1994), or graphical (Wickham, Cook, Hofmann, & Buja, 2010).

Second, Geiß (2021) suggests considering Q, n, and ρ in conjunction. In the realm of automated content analysis, there comes also ζ. Geiß (2021) suggest forming expectations regarding ρ and Q in design. Similarly, one should also form expectation regarding ζ. There is a great uncertainty in estimating ζ before the study because it depends on many factors: availability of training and validating data, the technology used to create the predictive models, the availability of computer resources etc. The website paperswithcode.com lists the state-of-the-art ζ of machine learning models. But in practice, we rarely can attain those levels (see Van Atteveldt et al., 2021 for examples).

Third, Geiß (2021) suggest preregistering the design together with these expectations. I concur.

2. Report cross tabulations, not just metrics

Metrics such as F1, CCR (for ζ) and Krippendorff's α (for Q) are useful for guiding one's methodological decisions or for comparing methods. For ζ, metrics such as F1 make assumptions about how ζ_+ and ζ_+ are averaged. F1, for example, has been criticized for generating misleading conclusions when the data are imbalanced (Chicco & Jurman, 2020). In the situation of $P(X = 1)$ being too high, a model with no talent in classifying true negative cases can still have a high F1, according to the equation (3).

Figure 4 displays the relationship between ρ and the observed $P(X = 1)$. In general, $P(X = 1)$ increases with ρ. It is because $P(X = 1|Y = 1)$ needs to be larger than $P(X = 1|Y = 0)$ in order to obtain a large ρ. Therefore, when ρ gets larger, the bias of F1 against true negative cases also gets larger. In those cases, the model does not need to know how to know how to classify true negative cases in order to get a high F1. Unlike F1,

the amount of random measurement errors generated by a model is measured by both the misclassifications of true negative cases and true positive cases. Figure 5 displays the simulation results of fixed ρ and n but with varying levels of ζ_+ and ζ_-. A model needs to have both high ζ_+ and ζ_- to obtain a relatively unbiased estimation of effect size. In medical research, it is common to report both ζ_+ (true positive rate, also known as sensitivity or recall) and ζ_- (true negative rate, also known as specificity). Reporting of prevalence-neutral measurements such as positive and negative likelihood ratios is also common (Hayden & Brown, 1999).

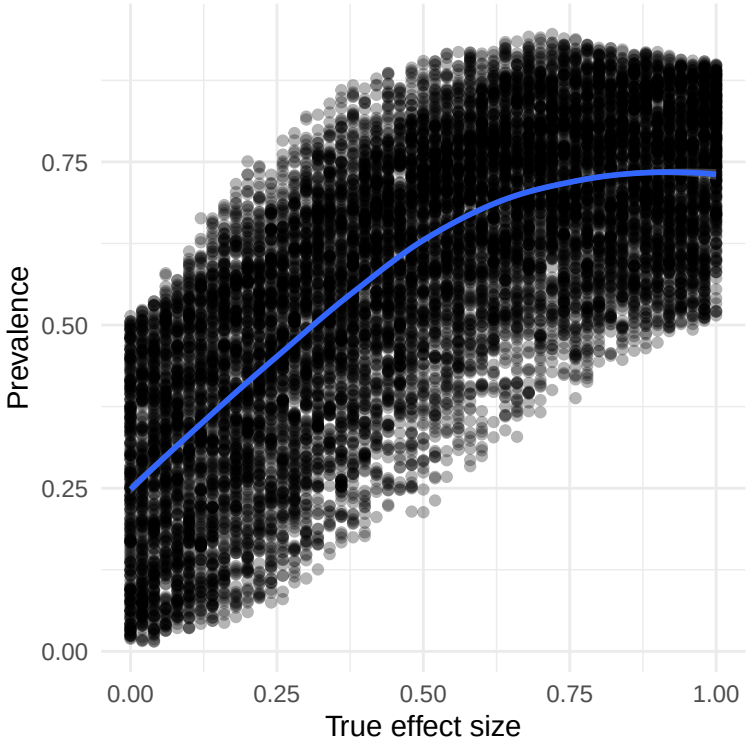

Figure 4. The relationship between true effect size and observed prevalence of X. The line represents the local regression line

Also drawing from medical research, the Standards for Reporting Diagnostic Accuracy (STARD) statement mandates the reporting of "Cross tabulation of the index test results (or their distribution) by the results of the reference standard" and "Estimates of diagnostic accuracy and their precision" (Bossuyt et al., 2015). Therefore, a medical

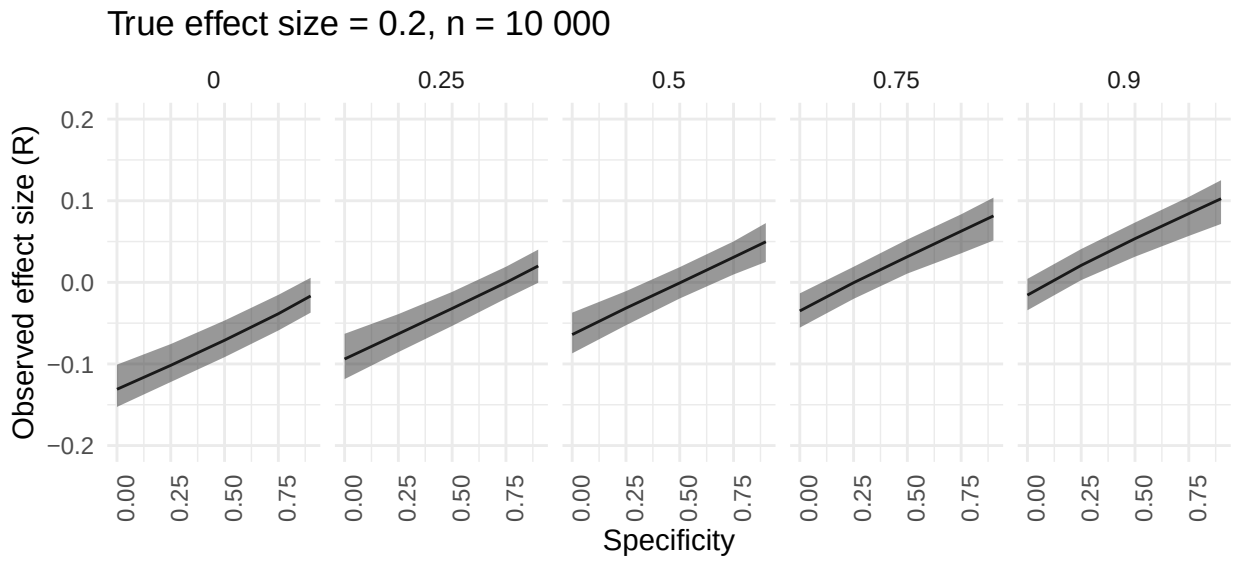

Figure 5. The relationship between true negative ratio (X-axis) and observed effect size (Y-axis) by true positive rate (column). The line represents the median of all simulated observed effect sizes and the ribbon represents its 95% interval estimation.

diagnostic study needs to report accuracy metrics, their precision (e.g. 95% Confidence Interval), and cross tabulations (See this paper for an example: Prince-Guerra et al., 2021). This practice should be ported to communication research.

The cross tabulation, which is called Confusion Matrix in machine learning literature, complements metrics. It is also more useful than the so-called heatmap visualization of Confusion Matrix because cross tabulation provides raw numbers, not just color depths. The numbers are important because they provide clue on whether the predictive accuracy data are subjected to random sampling error. A measurement of ζ calculated with only a few cases is less accurate than the ones with a lot of cases. The provided R function `cal_bigphi` also supports these numbers (the parameters `tpn`, `tnn`, `Qn`, the denominators for calculating ζ_+, ζ_-, and Q respectively) by assuming probabilities such as ζ_+, ζ_-, and Q are drawn from a binomial distribution, $\mathcal{B}(n, p)$. When these probabilities are calculated with a larger sample size, there are more accurate and subjected to less random sampling error.

```
cal_bigphi(rho = 0.0852, n = 2019,
           tpr = 0.831, tnr = 0.721,
           Q = 0.8964, tpn = 1000,
           tnn = 300, Qn = 400)
```

Regarding the number of cases for study Q and ζ, Song et al. (2020) advise researchers to "strive to increase the sizes of manually coded validation dataset as large as possible, preferably to more than N = 1,300 ..., assuming acceptable reliability (equal to or higher than .7)". I agree with this advice in principle but it is also important to get enough positive and negative cases so that both ζ_+ and ζ_- are accurate. However, this cannot be achieved by deliberately oversampling positive or negative cases, because of the problem below.

3. Rule out systematic measurement error

It is also important to note that the subject matter in this paper is *random* measurement errors. And all the simulation conducted in this paper assumes that *systematic* measurement error is negligible. As stated in the Introduction section, systematic measurement error could influence both the variance and the average level of a variable, but the direction of influence depends on the composition of data.

Unlike the Monte Carlo simulation, systematic measurement error exists in real life and it's not common for (communication) researchers to report it. In a methodological research, Song et al. (2020) study one source of systematic measurement error: the non-random selection of cases for validating prediction models and such practice will generate bias in the estimation of ζ. Therefore, it is inappropriate to oversample positive or negative cases to achieve higher n in the validation procedure. $P(X)$ should be kept like the natural distribution in the entire dataset. Therefore, the valid way to increase the number of positive and negative cases is to increase the number of random samples, not

deliberately select more positive or negative cases.

The composition of the cases for conducting the validation procedure is not the only source of systematic measurement error. Sometimes the error is not straightforward to detect and needs to be detected qualitatively. I recommend the error analysis process as in Van Atteveldt et al. (2021) to qualitatively document what kind of cases got misclassified.

Coda

The entry point for automated content analysis is to scale up content analysis for the ever-increasing size of datasets (Lewis, Zamith, & Hermida, 2013; Trilling & Jonkman, 2018). In other words, automated content analysis allows researchers to *ob*tain more statistical power through increasing n and to study some extremely small effect sizes without the risk of Type II error. The findings from this study, however, underscores the problem in *re*tain those obtained statistical power due to the automation-induced measurement errors. When we do not look at the problem in terms of statistical power, which is only relevant to the frequentist inference, the ACIDIC problem underestimates the estimated effect sizes.

Obtaining more statistical power and obtaining unbiased effect sizes are like weddings. One cannot dance at two weddings at the same time.

References

Bachl, M., & Scharkow, M. (2017). Correcting measurement error in content analysis. *Communication Methods and Measures*, *11*(2), 87–104. https://doi.org/10.1080/19312458.2017.1305103

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2020). Automated text classification of news articles: A practical guide. *Political Analysis*, 1–24. https://doi.org/10.1017/pan.2020.8

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., … Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, h5527. https://doi.org/10.1136/bmj.h5527

Chan, C.-h., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., … Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, 1–21. https://doi.org/10.1080/19312458.2020.1812555

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Corliss, G. (1977). Which root does the bisection algorithm find? *SIAM Review*, *19*(2), 325–327. https://doi.org/10.1137/1019044

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. Government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Dobbrick, T., Jakob, J., Chan, C.-H., & Wessler, H. (2021). Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. *Communication Methods and Measures*, 1–18. https://doi.org/10.1080/19312458.2021.1999913

Ethayarajh, K., & Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. *arXiv Preprint arXiv:2009.13888*.

Frost, C., & Thompson, S. G. (2000). Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *163*(2), 173–189.

https://doi.org/10.1111/1467-985x.00164

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Geiß, S. (2021). Statistical power in content analysis designs. *Computational Communication Research*, *3*(1), 61–89. https://doi.org/10.5117/CCR2021.1.003.GEIß

Gerring, J. (2012). Mere description. *British Journal of Political Science*, *42*(4), 721–746. https://doi.org/10.1017/s0007123412000130

Gilardi, F., Gessler, T., Kubli, M., & Müller, S. (2021). Social media and political agenda setting. *Political Communication*, *39*(1), 39–60. https://doi.org/10.1080/10584609.2021.1910390

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. https://doi.org/10.1177/0002716215569192

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Guo, L., Mays, K., Lai, S., Jalal, M., Ishwar, P., & Betke, M. (2019). Accurate, fast, but not always cheap: Evaluating "crowdcoding" as an alternative approach to analyze social media data. *Journalism & Mass Communication Quarterly*, *97*(3), 811–834. https://doi.org/10.1177/1077699019891437

Hayden, S. R., & Brown, M. D. (1999). Likelihood ratio: A powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. *Annals of Emergency Medicine*, *33*(5), 575–580. https://doi.org/10.1016/s0196-0644(99)70346-x

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *BMJ*, *340*(jun23 2), c2289–c2289. https://doi.org/10.1136/bmj.c2289

Knox, D., Lucas, C., & Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, *25*(1). https://doi.org/10.1146/annurev-polisci-051120-111443

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures*, *5*(2), 93–112. https://doi.org/10.1080/19312458.2011.568376

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting &Amp; Electronic Media*, *57*(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

Osborne, J. W. (2002). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. https://doi.org/10.7275/0K9H-TQ64

Prince-Guerra, J. L., Almendares, O., Nolen, L. D., Gunn, J. K. L., Dale, A. P., Buono, S. A., … Bower, W. A. (2021). Evaluation of Abbott BinaxNOW Rapid Antigen Test for SARS-CoV-2 Infection at Two Community-Based Testing Sites — Pima County, Arizona, November 3–17, 2020. *MMWR. Morbidity and Mortality Weekly Report*,

70(3), 100–105. https://doi.org/10.15585/mmwr.mm7003e3

Rains, S. A., Levine, T. R., & Weber, R. (2018). Sixty years of quantitative communication research summarized: Lessons from 149 meta-analyses. *Annals of the International Communication Association*, *42*(2), 105–124. https://doi.org/10.1080/23808985.2018.1446350

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv Preprint arXiv:2111.15366*.

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., … Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the facebook pages of news outlets. *New Media &Amp; Society*, *20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205

Trilling, D., & Jonkman, J. G. F. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*(2-3), 158–174. https://doi.org/10.1080/19312458.2018.1447655

Upton, G., & Cook, I. (2014). *A dictionary of statistics 3e*. Oxford university press.

Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation:

Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van Atteveldt, W., Velden, M. A. C. G. van der, & Boukes, M. (2021). The validity of sentiment analysis:Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 1–20. https://doi.org/10.1080/19312458.2020.1869198

Western, B., & Jackman, S. (1994). Bayesian inference for comparative research. *American Political Science Review*, *88*(2), 412–423. https://doi.org/10.2307/2944713

Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 973–979. https://doi.org/10.1109/tvcg.2010.161

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *75*(6), 579. https://doi.org/10.2307/2340126