Summer 7-13-2020

Assessing the Validity of Sentiment Analysis Measures through Polychoric Correlation

Kelli N. Kasper
University of New Mexico

Kelli Nicole Kasper

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Dr. Fletcher G.W. Christensen , Chairperson

Dr. Erik B. Erhardt

Dr. Ronald Christensen

Assessing the Validity of Sentiment Analysis Measures through the Use of Polychoric Correlation

by

Kelli N Kasper

B.S. Statistics, UNM, 2017
B.S. Psychology, UNM, 2017

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Statistics

The University of New Mexico

Albuquerque, New Mexico

July, 2020

Dedication

To my mom, the most mean mom there ever was. My dad, who reminded me the whole time that he almost failed Introduction to Statistics. And to my partner, who always believed and pushed me to the very best I could ever be.

Acknowledgments

Mister Chad B. Kasper, who was the inspiration for doing all of this.

My advisor, Dr. Fletcher G.W. Christensen, for reminding me that you will always be cooler with two middle initials.

My entire family, who had to wonder what was taking so long.

Grace Mayer, who started the experience of sentiment analysis with me.

My fellow graduate students, who knew why it took so long.

Tristan E. Jerome, for always bringing me a cup of tea and biscuit when I looked most stressed.

Assessing the Validity of Sentiment Analysis Measures through the Use of Polychoric Correlation

by

Kelli N Kasper

B.S. Statistics, UNM, 2017
B.S. Psychology, UNM, 2017

M.S., Statistics, University of New Mexico, 2020

Abstract

Sentiment analysis methods extract the attitude of a text via systematic algorithms. To evaluate the validity of common sentiment analysis methods, we use polychoric correlation to compare computer-mediated methods and human-rated analogues. Our main topics of interest are the internal consistency of the raters' scores, the level of consensus among raters, and how well raters' scores correlate with those given by sentiment analysis methods for randomly collected Twitter data.

Our analysis found that there is good validity for methods that measure negative and positive sentiments in short texts, both in terms of inter-rater consistency and when comparing raters to computer-mediated sentiment analysis methods. The more complex sentiment pair anger and joy had lower levels of inter-rater consistency. Rater-computer consistency for anger and joy was the weakest among the methods tested, raising questions about the construct validity for measuring expressions of anger and joy in short texts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter we will introduce sentiment analysis: what it is and where it is useful, previous methods of examining and validating sentiment analysis techniques and why polychoric correlation is a useful metric for the validation of sentiment analysis constructs.

1.1 Sentiment Analysis, Background

Everyone is a content creator or influencer. Everyone is a critic or a self-proclaimed expert. Many people want to contribute their two cents, and everyone wants the world to know what they think about anything at any time. With every enter, submit, and publish, the internet is generating more data than can ever be analysed and used—but not for lack of trying. The mass of information that is generated has given rise to methods for analyzing this "big" data. Among those new tools, sentiment analysis addresses the vast amount textual data that is created every time a president gives a speech, a journalist publishes their daily blog, or a graduate student posts a tweet.

Sentiment analysis (SA) refers to a collection of methods for extracting the attitude of a text via systematic algorithms. SA methods quickly summarize text or

texts, often in either an unsupervised or semi-supervised system. Similar to other types of mass data analysis, sentiment analysis quickly assesses "big" data sets of textual data, of which there is a continuously growing supply. These methods provide tools for addressing many analytical goals. Three common examples include: gauging the political atmosphere pertaining to a particular topic, sorting product reviews, and comparing styles of different authors. The examples below demonstrate the wide range of research areas where SA methods are used.

In a 2011 article, Leetaru used sentiment analysis to examine tone in several news outlets. Using the New York Times (NYT) and looking at the average monthly tone from 1945 to 2005, he found that NYT has become increasingly negative since the 1960s (Leetaru, 2011). Pinker used this analysis to question why the news has become more negative when almost all standards of life, health, literacy, and safety have measurably improved in recent decades (Pinker, 2018). Sentiment analysis provided a metric to quantify the negativity of the news that could be contrasted with the many metrics of progress in Pinker's work on news pessimism and how it impacts consumers.

At the University of Virginia, Chen et al. (2015) combined sentiment analysis with GPS-tagged Twitter posts (tweets) and weather predictors to improve on a kernel density estimation of crime prevalence using historical crime data. Using a lexicon-based SA method to determine the polarity of tweets, Chen et al. constructed a linear model to predict the time and location where specific types of crimes would occur and found their model outperformed the previous model that used only kernel density estimation.

And recently, Weissman et al. (2019) employed six sentiment analysis methods to evaluate the subjective expressions in clinician's hospital notes of critically ill patients. They worked to assess the construct validity of existing sentiment analysis methods and use it to improve on early predictions of in-hospital mortality. Although unable to improve on current methods of predicting in-hospital mortality, they did find that positive sentiments measured by several of the methods—including the Bing and NRC lexicons discussed in Chapter 2—were inversely related to death on

the same day, after adjustments for demographics and severity of illness. Weissman et al. remain convinced of the usefulness of sentiment analysis for the health field, but recommended that a discipline specific lexicon be used for improved results.

These uses of sentiment analysis provide an idea of the range of its applications. Twitter data, in particular, is a promising resource for sentiment analysis studies. The sheer volume of textual information created on Twitter, coupled with the fact that tweets often contain time and location information, make tweets a data-rich source for studies like Chen et al. Tweets have a tight character limit, however—140 characters before Nov. Nov. 2017 and 280 characters since then. Because the limited amount of textual information contained in tweets may decrease their legibility to sentiment analysis methods, we undertake to examine the construct validity of sentiment analysis methods when applied to data of this type.

1.2 Validity of Sentiment Analysis Tools

The concept of test validity is a mid-century innovation in the field of psychometrics (Cronbach and Meehl, 1955). Validity refers to the degree to which a measurement method reflects what it purports to measure. Sentiment analyses involve the attribution of some sentiment to a corpus of text—sometimes as simple as saying that a text is positive or negative, sometimes indicating that a text exhibits a particular emotion like joy or anger (Haseena Rahmath and Ahmad, 2014). Sentiment analysis methods use various operational definitions of the sentiments they investigate, for example by considering a curated lexicon of words that are thought to express the sentiment of interest. But we should not understand the operational definition used by the sentiment analysis measure as, necessarily, an operational definition of the sentiment itself. To give a simple example, if a lexicon-based analysis method were to assess "bad" as a word indicating negative sentiment, then it might assess as negative the phrase, "Not bad. Not bad at all." We are, therefore, interested in the extent to which the assessments of sentiment analysis methods agree with the constructs they are supposed to measure.

Cronbach and Meehl (1955) explain the differences between various types of validity assessments. We will concern ourselves with two particular notions of validity: criterion validity and contstruct validity. Cronbach & Meehl define construct validation as, "[The process] involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.' " Criterion validation, conversely, assesses the extent to which results from a test accord with some other criterion that is accepted to be measuring the thing that is desired. Negative sentiment, positive sentiment, anger, joy, etc., do not have clear operational definitions—but their semantic definitions may be strong enough to allow us to assess validity of sentiment analysis methods by comparing their results to naturalistic evaluations of semantic content done by human raters.

Before we consider these comparisons of evaluations, though, we consider the work that has already been done to assess the reliability of sentiment analysis.

1.2.1 Three Tests of Reliability

In a comparison study by Chalothorn and Ellman (2012), SentiWordNet and SentiStrength (two lexicon-baseds SA methods detailed further in Chapter 2) were both used to examine the negativity and positivity of web forms. To compare the different methods of sentiment analysis, the authors used SentiWordNet and SentiStrength to classify posts as either positive, negative or neutral. From those classifications, the overall percent of each classification for each of the methods was compared to one another. The first form was classified to have 35% negative posts when using the lexicon SentiWordNet and 50% of the posts were classified as negative when using SentiStrenth. The second form was classified to have 15% negative posts when using the lexicon SentiwordNet and 30% negative posts when using SentiStrength.

In a meta-analysis of the concepts of opinion mining and sentiment analysis, Haseena Rahmath and Ahmad (2014) examine multiple methods of sentiment analysis from published results. They report the approach, data set, and technique used; as well as the accuracy of the technique. Accuracy is described as the percent of

the time that the method used was correct in comparison to a predetermined categorization. They found that lexicon-based sentiment analysis methods had between a 71% to 82.4% accuracy and that some lexicon and machine learning hybrid methods had between a 66.8% to 85.4% accuracy. They concluded that using lexicons was a simpler method that did not require supervision or training making it useful for faster results; machine learning methods, although performing better in general, needed large labeled training data sets to be effective.

A comparison study of lexicon-based and machine learning sentiment analysis methods used Cornell's Movie Review Data: two thousand movie blogs (reviews) as a data set, half of which were classified before the study as positive and the other half as negative. The classification process of the movie blogs used a system of ad-hoc rules to determine if a blog as positive or negative. Primarily, this was done by using the number of stars or letter grade given in the review, with a lower number of stars or low letter grade indicating a negative review and visa versa (Pang and Lee, 2004). The authors measured the accuracy as the percent of trials that the examined process matched the initial described sentiment of the blog for each method (Annett and Kondrak, 2008). The accuracy of each methods and any variations upon them were then compared. When using lexicon based sentiment analysis, they had accuracy between 50% and 60.4%. Machine learning SA method accuracy was found to be between 65.4% and 77.5%.

Two of these previously described methods of testing the validity of sentiment analyses depend on the same metric, accuracy, while the other simply compares the percent of classification from one method to the other. While these are useful ways of comparison there is more that can be done to show the construct validity of these methods. The validation of sentiment analysis methods can be expanded beyond whether or not it is correctly classified by both a preset measure and the technique, and there is more to do in comparing sentiment analysis to individually rated pieces of text and even the degree of sentiment in each rating.

1.3 Proposal to Use Polychoric Correlation

More than a simple judgment of whether a sentiment is present or absent, SA methods often return scores on a scale from low sentiment to high sentiment with ordinal values assigned to describe the level of sentiment present. This stratification adds more granularity to the rating system than a simple judgment of whether a sentiment is present or absent, which is more akin to how people might describe sentiment. This presents new opportunities and challenges in comparing and contrasting methods.

Ordinally ranked variables are not truly quantitative, however they can be a simplified way to summarize a continuous variable into a few discrete levels. From this latent continuous property of the variables, polychoric correlation can be used to describe the relationship between ordinal variables with the same interpretation as the traditional Pearson product-moment correlation.

Polychoric correlation is an extension of tetrachoric, correlation first developed by Pearson to provide a measure of association of the variables from a two-by-two contingency table. The method assumes that there is a latent bivariate normal distribution being represented in the dichotomous framework of the contingency table. In tetrachoric correlation, some threshold divides the continuous spectrum into two parts which are recorded for each variable. Polychoric correlation is an extension of the original idea of bivariate normal distributions underlying joint ordinal responses—but now each ordinal response represents some spectrum of the continuous latent variable, with R and S possible responses for each variable, and $R-1$ and $S-1$ thresholds dividing the spectrum into the observed levels.

Rater reliability in testing is a common concern in multiple disciplines such as physiology and psychology where, from rating to rating, a participant's scores can be inconsistent. Polychoric correlation provides a useful method to validate a raters reliability as well as contribute to the construct validation of the method when scores are given in polytomized form. Polychoric correlation quantifies the marginal homogeneity between two different raters to determine consistency or a lack of it. Similarly it provides a quantitative association between two separate category definitions'

agreement in evaluating, even when the definitions vary in the number of levels used. Values of polychoric correlation close to zero would indicate little agreement between raters or measurement scores and values closer to 1 would indicate a higher level of agreement. A polychoric correlation close to -1 could indicate that the raters are either in complete disagreement, or that their scales have reversed scoring (i.e. one scale might rate a high level of depression as a 1, while another scale may rate its highest level of depression as a 10). High correlations and high stability can be used as evidences to the construct validity of a measurement (Cronbach and Meehl, 1955) and using polychoric correlation as a proxy for Pearson product-moment correlation, we can assess the validity of the sentiment analysis methods investigated in this study with more precision than previous studies have managed.

1.4 Polychoric Correlation for Sentiment Analysis

The term *sentiment analysis* encompasses two different types of methods to perform interpretations of text. The first type is the lexicon-based method which uses corpora, lists of terms, to identify key sentiment terms. The second type is machine learning which uses algorithms (Haseena Rahmath and Ahmad, 2014) to make decisions based on training data, which then informs the analysis of sentiment on a text. All of these methods are an attempt to generate a process that can quickly be applied to many thousands of texts, books, essays, or speeches to emulate a human evaluation of sentiment type and strength. However, none are perfectly analogous to what a person may naturalistically determine the sentiment of a text to be. New lexicons with specific topics are created, different machine learning algorithms are designed, all to provide a more precise estimate of human judgment. In this aspect, polychoric correlation can be used to compare human raters' decisions to these automated sentiment analysis tools.

Most sentiment analysis techniques report a single aspect of sentiment at a time such as negativity, positivity, joy or anger. Therefore, to compare with the results of a particular sentiment analysis, participants can read a text and rate it with what they

feel to be an appropriate score using a Likert scale. This type of scale is a common form of measurement in which scores range over a small range of values. Likert scales are frequently used to measure agreement with a text, where the options may be: "Agree strongly", "Agree", "Neither agree nor disagree", "Disagree", and "Disagree strongly", for example, with each option also being given a numeric equivalent. We can also imagine a Likert scale for a sentiment running from 0 to 5, with a zero indicating that the text contained none of the sentiment that is being investigated, a one indicating that there is low amount of sentiment, a three indicating a moderate amount of sentiment, and a five indicating a high level of sentiment. If more than one sentiment is being evaluated on the same text, randomizing the text and current sentiment could help reduce a carry-over bias from evaluating the same text several times in a row with varying sentiments. By comparing SA method results with participants' scores, the validity of the sentiment analysis tools can be evaluated.

There is an additional advantage to using polychoric correlation to compare the results of sentiment analysis when results are given as an ordinal response. In a number of the applications described previously, accuracy or success was defined as two methods providing the same classification. With such inflexibility there is a potential to see less "success" than there might actually be. As polychoric correlation describes the latent linear relationship between the two variables, it provides a more flexible method to compare results.

Chapter 2

Methods

2.1 The Data

Twitter consists of micro-blog posts known as tweets. Due to their restricted length, 280 characters or less, most are concise and focused on a singular subject. For this assessment the concise nature of tweets is ideal for allowing raters to look at many sentiment-dense examples of text for comparison.

The data used for the purposes of this study came from a self-collected random sample of social media posts on Twitter, collected between April 12, 2019 and February 8, 2020. Collecting data over a long span of time, as opposed to a shorter period, helps reduce the probability of using data that shares an underlying condition which might cause the contents of tweets to share topics and potentially induce rater bias.

Prior to data collection, a Twitter Developer Account was established with permissions to collect data for academic use. To collect tweets, a computer with the application "Task Scheduler" was programmed to execute a script, which is further detailed in Section A.1 of the appendix. This script collected and saved a sample of 500 tweets based on the search term " " (a space), from accounts with declared language English, every five minutes for a duration of the collection period. Using a space as a search term was determined to be the least biasing search term that also

ensured that the presence of some text content in the tweet. The decision to collect 500 tweets every five minutes was based on (a) the limitations of the Developer Account, which was only permitted to download 1500 tweets every fifteen minutes; and (b) the Task Scheduler application, for which five minutes was the smallest timing option. Additionally, to prevent files from growing too large, between data collections the current data file was periodically renamed and a new null file created to replace it, into which new tweets would be appended.

During the process of data collection, two data files—containing dates from October 9th, 2019 to October 29th, 2019 and November 23rd 2019 to December 13th, 2019—were corrupted. Data from these times were inaccessible, and thus not included in our sampling frame. Additionally, due to periodic internet connectivity problems, a full set of data was not collected for some days. Days with significantly less than full data were excluded from the final sampling frame. Both of these missing data issues are considered missing completely at random (there is no reason to think the missingness mechanism would relate to the topics we're trying to study) and we determined that missingness was unlikely to have biased the data.

From the full sampling frame of Twitter social media posts, 10 random days were selected to perform sentiment analysis on, creating a stratified sample. To ensure a variety of levels of sentiment for raters to assess, a sample of 5 tweets from each day were chosen systematically by using the results of the sentiment analysis methods under study. This systematic subsampling used the scores of the SA methods for sentiments: negative, positive, anger and joy. Because the negative and positive sentiments had 5 corresponding scoring systems the data was divided into two levels, high and low, corresponding to the five different rating levels. Four tweets were selected that had sentiments corresponding to 10 or less on AFINN scoring and 4 or less on Bing, NRC, SentiWordNet and SentiStrength scoring. Then one tweet was selected with scores above the previously described scores, for each of the ten days. This process was done for both the negative and positive sentiment scores to provide an overall sample size of 100 tweets: 40 lower negative scoring, 40 lower positive scoring, 10 higher negative scoring, and 10 higher positive scoring.

As the sentiments anger and joy only appear in one lexicon, NRC, we were able to use a more diversified systematic process for these sentiments. One tweet was selected that contained sentiment equal to zero, the next tweet was selected with sentiment scores of 1 or 2, two tweets with sentiment equal to 3 or 4 were selected and finally a tweet with sentiment greater than or equal to 5 was selected. This was done for both the anger and joy sentiments and the sample was combined to make a second sample of 100 tweets with a variety of sentiment scores.

This process of obtaining a wider range of tweets ensured that there was a greater amount of variability in sentiment to be rated by our human raters. The increased variability makes it easier to see the relationship between the computer generated sentiment scores and human rated scores on the same tweets.

For readability, tweets were put through a cleaning process that removed hyperlinks and some web page coding symbols that would disturb the content of the tweet to be examined. However, to keep tweets consistent with a daily reading style for the raters to assess, stop words were not removed. Stop words (e.g. a, the, that) are very common but non-sentiment-related words that are typically removed because they make the process of sentiment analysis take longer computationally. Since our purpose is to compare human scores with the scores produced by computer-mediated sentiment analysis methods, the decision to leave stop words intact was made.

2.1.1 Human Raters

Using a factorial design, eight volunteers were randomly assigned to one of eight treatments. These human raters then provided scores on three sentiment scales created to mimic the process of the five computer-mediated sentiment analysis methods we will study, to allow for more direct comparison with those methods. To control for as much error as reasonably possible, the randomized factorial design was comprised of the following factors:

- type of paired sentiment examined (negative/positive or anger/joy),

- order of which sentiment was rated first,
- whether the rater received a training document that gave them priming words to better help them identify the target sentiment, or a neutral training document without priming words.

The training documents used and the process for choose priming words are detailed in Section A.3 of the appendix.

The volunteers were given their first training document related to the first sentiment, and then read and scored the 100 tweets of their assigned sentiment sample. Raters gave each tweet scores according to three scales, which correspond to the different scoring systems of the computer-mediated sentiment analysis methods. After the first set of ratings, a fifteen-minute break was taken where volunteers were asked to relax and not to do any activity that could be linked to feelings of high sentiment, such as going on social media or checking the news. Avoiding potentially sentiment-priming activities was done to reduce the amount of bias raters may have while rating. Engaging in simple tasks or games such as Tetris was encouraged. The goal of this break was to give raters time to refresh themselves from considering the first sentiment, without biasing themselves before rating the next. After the break, volunteers were given their second training document and asked to rate the tweets again. The order of the tweets was randomized for every trial.

2.1.2 Scoring Process

As previously discussed, raters were asked to score their selection of tweets on three scales. For scale 1, raters gave a score of 0–5 to rate the overall sentiment of the text. For scale 2, they counted the number of words with sentiment in the text and reported the total number of such words as their score. Finally, on scale 3, they assigned a value of 1–5 for each word of sentiment, as previously identified in the second scale, and added those scores to obtain a final rating. The training documents and details of the scoring are given in Appendix A.3.

Each of these three rater scales parallels one or more of the discussed computer-mediated tools for sentiment analysis. Scale 1 parallels the construction for SentiStrength. Scale 2 parallels the construction for three different sentiment analysis measures: Bing, NRC and SentiWordNet. Finally Scale 3 parallels how the AFINN measure is constructed.

2.2 Description of Sentiment Analysis Tools

For our comparisons, five commonly used lexicon-based methods of sentiment analysis were used. Each method either had a unique method of constructing the corpus, or a different style of analysis. Lexicon-based sentiment analysis methods break a text into token words, and then compare these to a lexicon of words that have been judged to exhibit a target sentiment in order to create a sentiment score for the text.

2.2.1 AFINN

The AFINN-111 lexicon created by Finn Arup Nielsen, and commonly referred to as AFINN, is comprised of a list containing 2477 English words and phrases (Nielsen, 2011). The AFINN lexicon was designed to consider micro-blog posts, such as tweets, and uses a dictionary-based approach to compare text with its list of classified words. Given the restricted nature of Twitter posts, a word-matching scheme such as the AFINN would be useful to detect sentiment in short, informal and direct lines of text. The AFINN lexicon is one of the most popular methods to use in sentiment analysis, especially when examining Twitter (Koto and Adriani, 2015).

Relative to the other lexicons in this study, AFINN draws upon a smaller list of reference terms. However, each term has been scored with values ranging from negative five to positive five. A more negative score indicates stronger negative sentiment

and a more positive score indicates stronger positive sentiment, while values closer to zero are to be considered neutral. This type of scoring allows for detection of impactful language that could be otherwise overlooked. The AFINN lexicon is available through the statistical analysis program R and the package TidyText (Silge and Robinson, 2017), making it a popular and freely available option for sentiment analysis.

2.2.2 Bing

The Bing lexicon (named for the author not the search engine) was developed by Professor Bing Liu of the University of Chicago and his research team. Considered a large contributor to the field of sentiment analysis, Liu (2011) wrote the book *Opinion Mining and Sentiment Analysis*, and has authored and coauthored many articles about sentiment analysis and opinion mining. Liu is considered among the front-runners of sentiment analysis, being named as a fellow of the Association for Computing Machinery and Association for the Advancement of Artificial Intelligence for his contributions to opinion mining and sentiment analysis . Initially, this lexicon was developed to determine opinions in product reviews, but is highly recommended as a starting point for other applications of sentiment analysis (Silge and Robinson, 2017).

The Bing lexicon takes a binary dictionary approach where terms are considered either positive or negative. After determining the word polarity, this lexicon method can be used to display the count of either in a selected text. 6,788 terms have been categorized as either positive or negative for this manually created lexicon. Additionally for this lexicon, terms are not limited to words; word-number hybrids and misspelled words are included as well, making it particularly useful for Twitter data where words are often shortened, misspelled and abbreviated with numbers to save character space. The Bing lexicon was given express permission to be used in the R program TidyText (Silge and Robinson, 2017).

2.2.3 NRC

The NRC or Word-Emotion Association lexicon was created by Saif Mohammad and Peter Turney (Mohammad and Turney, 2013). Another manually created lexicon, the construction of this lexicon differs in that the developers used the "combined strength and wisdom of the crowds" (Mohammad and Turney, 2013) to make a high quality, word emotion associated lexicon. This was done by asking participants if a term evoked a particular emotion. This type of construction of a lexicon that has been made by the masses aids itself to the analysis of social media such as Twitter where millions of people post daily.

Containing 13,901 terms, this lexicon has classified words as expressing the following sentiments: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The wide variety of sentiments available in the NRC lexicon give researchers a variety of emotions to look for and focus on for any particular analysis. This lexicon can be reduced down to positive and negative, to make it more comparable to the other SA methods discussed, and used to search tweets and count the occurrences of those words. With 3,324 negative terms and 2,302 positive terms, the reduced negative-positive (NP) NRC lexicon is comparable in size to previously mentioned lexicons. NRC is one of the few lexicons also accessible through TidyText (Silge and Robinson, 2017), making it easily accessible and recommended as a tool for sentiment analysis (Welbers et al., 2017).

2.2.4 SentiStrength

One of the most common baseline measures of sentiment is SentiStrength, a method that is almost always used as a comparison for new techniques and lexicons (Thelwall, 2017; Nielsen, 2011) and is used by most studies in sentiment analysis (Jongeling et al., 2015). SentiStrength was specifically used for the AFINN lexicon (Nielsen, 2011) for comparison.

SentiStrength is an independent software that performs sentiment analysis using

a corpus-based lexicon. Taking into account preceding words, SentiStrength looks at the sentiment of words and phrases including negations such as "not good," which it would correctly categorize as negative. SentiStrength also contains rules about punctuation and misspellings to help determine sentiment. Initially trained and built using MySpace comments (Thelwall et al., 2010), SentiStrength is well suited for social media analysis and is still highly used for analyzing text from many social media platforms such as YouTube comments (Mäntylä et al., 2018). Recently an entire chapter of *Sentiment Strength Detection in Short Informal Text* (Thelwall, 2017) was dedicated to how to use SentiStrength. SentiStrength can output sentiment scores either in a binary fashion of -1 or 1 for negative or positive sentiment, or on a scale from -4 to 4. SentiStrength is freely available for research and can be downloaded from

`http://sentistrength.wlv.ac.uk/`

2.2.5 SentiWordNet

The last computer-mediated sentiment analysis method we consider is SentiWordNet, created by Esuli and Sebastiani (2006) and Baccianella et al. (2010). SentiWordNet uses a lexicon created using a semi-supervised learning algorithm to choose terms specifically useful for sentiment analysis. Starting with a small group of words that were classified as either positive or negative, the semi-supervised learning algorithm extended the list of words by finding similar terms and classifying them the same as previously listed words. Between the first and third edition of SentiWordNet this method was used in 2.7% of all searches related to sentiment analysis, making it the most commonly searched lexicon-based SA method (Mäntylä et al., 2018).

With one of the largest lexicons, SentiWordNet contains 20,093 terms categorized as either positive or negative. This extraordinarily large lexicon makes it useful for handling the large variety of potential language used on Twitter. A version of the lexicon is made available in the sentimentr and lexicon R packages by Tyler Rinkle.

2.3 Computing Polychoric Correlation through Maximum Likelihood

In this section, we detail the method of polychoric correlation. This discussion relies heavily on Olsson (1979) for its presentation of maximum likelihood methods for estimating these correlations.

First, let us observe two ordinal scales, represented with x and y. These are classified into categories $s \in \{1, ..., S\}$ and $r \in \{1, ..., R\}$. A cross-tabulation table of x by y gives the observed frequencies as denoted in the table below, where the frequencies are the number of times an observation had the pair of ordinal values x_s and y_r assigned to it. We assume (Olsson, 1979) x and y are discretizations of some latent variables ξ and η, which are bivariate normal distributed. We define thresholds $a_0, ..., a_S$ for partitioning ξ into x, and $b_0, ..., b_R$ for partitioning η into y. The relationship between x and ξ can be written as:

$$a_0 = -\infty$$

$$a_s = \infty$$

$$x = i \text{ if } a_{i-1} \leq \xi < a_i$$

Similarly the relationship between y and η can be written as:

$$b_0 = -\infty$$

$$b_r = \infty$$

$$y = j \text{ if } b_{j-1} \leq \eta < b_j$$

Now we consider a data set of n observations. Each observation can be expressed as (x_{ik}, y_{jk}) where $k \in (1, 2, .., n)$. As stated before, x has i levels where $i \in (1, 2, ..., s)$ and y has j levels were $j \in (1, 2, ..., r)$. We let n_{ij} be the frequency of occurrence for

the combination of levels i and j, so $n = \sum_{i,j} n_{ij}$. Further, we can give the table of cross-tabulations as follows:

	1	2	3	...	r
1	n_{11}	n_{12}	n_{13}		n_{1r}
2	n_{21}	n_{22}	n_{23}	...	n_{2r}
3	n_{31}	n_{32}	n_{33}	...	n_{3r}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s	n_{s1}	n_{s2}	n_{s3}	...	n_{sr}

Table 2.1: Observed frequencies

In Table 2.1 we have the discretized representation of a bivariate normal distribution. Using the latent bivariate normal relationship between x and y, the method described below uses the occurrence frequencies n_{ij} to estimate the thresholds needed for the computation of polychoric correlation.

2.3.1 Estimation of parameters

The parameters a_i and b_j are thresholds defining the discretization of ξ and η into x and y respectively. Our goal is to estimate correlation ρ between ξ and η, given the data in the form of the cross-tabulation table.

The maximum likelihood method is used to estimate ρ and the thresholds a_i and b_j simultaneously. Our data for doing this estimation consist of the observed frequencies n_{ij}. Let π_{ij} be the probability of an observation being categorized into cell (i, j). Then the likelihood of the sample is

$$L(\pi_{ij}|n_{ij}) = C \prod_{i=1}^{s} \prod_{j=1}^{r} \pi_{ij}^{n_{ij}},$$

where C is a constant.

Then, $\ell(\pi_{ij}|n_{ij}) = \ln(L(\pi_{ij}|n_{ij})) = \ln(C) + \sum_{i=1}^{s} \sum_{j=1}^{r} n_{ij} \ln(\pi_{ij})$.

Since the thresholds for x are denoted by a_i and y are denoted by b_j, and as both the latent variables ξ and η are assumed to be normally distributed, it follows that

$$\pi_{ij} = \Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) - \Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1}),$$

where Φ_2 is the bi-variate normal distribution function with correlation ρ, the latent correlation of interest. A more standard notation is $\Phi_{2,\rho}$, however, because the assumption of covariance ρ will hold everywhere we use this function, we abbreviate with Φ_2. This will greatly simplify notation to exclude the subscript everywhere since it can be assumed to apply everywhere.

To estimate the correlation, the parameters a_i, b_j, and ρ all need to simultaneously be estimated. This requires taking the partial derivative of the log likelihood with respect to each of these parameters, so we can find the roots of the full likelihood surface.

First we must find the first derivative of the log likelihood with respect to ρ:

$\frac{\partial}{\partial\rho} l(\pi_{ij}|n_{ij}) = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \rho}$

To find that value, we must consider π in terms of the latent bivariate normal random variables, so:

$$\begin{aligned} \frac{\partial}{\partial\rho}\pi_{ij} &= \frac{\partial}{\partial\rho}\Big[\Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) - \Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1})\Big] \\ &= \phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1}) \end{aligned}$$

(Johnson and Kotz, 1972).

Then,

$\frac{\partial}{\partial\rho} l(\pi_{ij}|n_{ij}) = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} [\phi_2(a_i, b_j) - \phi_2(a_{i-1}, b_j) - \phi_2(a_i, b_{j-1}) + \phi_2(a_{i-1}, b_{j-1})]$

Next is to find the first derivative of the log likelihood with respect to a_k.

$\frac{\partial}{\partial a_k} l(\pi_{ij}|n_{ij}) = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial a_k}$

We note that observations in n_{ij} only provide information about the thresholds immediately above and below their cell. That is, knowing that an observation of ξ translates into an observation in the n_{ij} cell indicates that $a_{i-1} \leq \xi < a_i$ and this observation provides no information about, say, a_{i-2} except through the information it provides about a_{i-1}. This means that our partial derivatives can be expressed as:

$$\frac{\partial \pi_{ij}}{\partial a_k} = \begin{cases} 0 & \text{if } i \neq k \ \& \ i \neq 1+k, \text{ i.e. } \pi_{ij} \text{ does not contain } a_k \\ \frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} & \text{if } k = i \\ -\frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} + \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} & \text{if } k = i-1 \end{cases} \tag{2.1}$$

This means that we need only let i go from k to $k+1$ in calculating the partial derivative of the log likelihood with respect to a_k. We can write that:

$$\begin{aligned} \frac{\partial}{\partial a_k} l(\pi_{ij}|n_{ij}) &= \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial a_k} \\ &= \sum_{j=1}^{r} \frac{n_{k,j}}{\pi_{k,j}} \Big[\frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} \Big] \\ &\quad + \frac{n_{k+1,j}}{\pi_{k+1,j}} \Big[- \frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} + \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} \Big] \\ &= \sum_{j=1}^{r} \Big(\frac{n_{k,j}}{\pi_{k,j}} - \frac{n_{k+1,j}}{\pi_{k+1,j}} \Big) \Big[\frac{\partial \Phi_2(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi_2(a_k, b_{j-1})}{\partial a_k} \Big] \end{aligned}$$

And as ϕ_1 and Φ_1 denote the univariate normal density and distribution functions, we can use the form:

$\frac{\partial \Phi_2(u,v)}{\partial u} = \phi_1(u) \times \Phi_1 \Big[\frac{(v - \rho u)}{(1-\rho^2)^{\frac{1}{2}}} \Big]$

(Johnson and Kotz, 1972).

Finally, the first order derivative of the log likelihood with respect to a_k can be found to be:

$$\frac{\partial}{\partial a_k} l(\pi_{ij}|n_{ij}) = \sum_{j=1}^{r} \left(\frac{n_{k,j}}{\pi_{k,j}} - \frac{n_{k+1,j}}{\pi_{k+1,j}} \right) \times \phi_1(a_k) \left[\Phi_1 \left(\frac{b_j - \rho a_k}{(1-\rho^2)^{\frac{1}{2}}} \right) - \Phi_1 \left(\frac{b_{j-1} - \rho a_k}{(1-\rho^2)^{\frac{1}{2}}} \right) \right]$$

And then by symmetry we can find the first order derivative of the log likelihood with respect to b_m to be:

$$\frac{\partial}{\partial b_m} l(\pi_{ij}|n_{ij}) = \sum_{i=1}^{s} \left(\frac{n_{i,m}}{\pi_{i,m}} - \frac{n_{i,m+1}}{\pi_{i,m+1}} \right) \times \phi_1(b_m) \left[\Phi_1 \left(\frac{a_i - \rho b_m}{(1-\rho^2)^{\frac{1}{2}}} \right) - \Phi_1 \left(\frac{a_{i-1} - \rho b_m}{(1-\rho^2)^{\frac{1}{2}}} \right) \right]$$

At this point we have the partial derivative formulas for ρ and all the thresholds. We can use these to find roots on the likelihood surface, with the intention of looking for choices of these parameters that maximize the likelihood. However, solving large systems of linear equations is difficult, so, we will move to easily available computational methods for estimating these values, as detailed in Section 2.3.2.

2.3.2 Using R to Calculate Polychroric Correlation

The package "polycor" by Fox (2007) provides a well documented and supported function, polychor to calculate the polychoric correlation by estimating the correlation coefficient ρ and the thresholds a_i and b_j simultaneously.

Using the function polychor the maximum likelihood estimator is computed by maximizing the bivariate normal likelihood with respect to the thresholds inducing the two discretized variables x and y, with r and s levels respectively. Then the polychoric correlation is then the simultaneous maximum likelihood estimate of ρ for the latent bivariate normal random variables, ξ and η. The likelihood is maximized numerically using the optim function, and the co-variance matrix of the estimated parameters is based on the numerical Hessian computed by optim function.

The values provided by the polychor function will be used to assess how well the different sentiment measures match up with one another, as a way of evaluating the validity of the underlying textual sentiment constructs.

Chapter 3

Results

With data from the five different sentiment analysis tools and eight raters we can inspect the validity of sentiment analysis. Our main topics of interest are: the internal consistency of the raters' scoring across the types of scores that were given, the level of consensus among raters, and how well raters scores correlate with those given by sentiment analysis methods. These analyses will let us assess the validity of the included sentiment analysis constructs, are these methods representative of what raters measure when they're asked to evaluate the sentimental content of tweets themselves. Secondarily, we have included several additional analyses to address if raters can consistently rate sentiment without systematic biases in their rating including: order of sentiment rating, training, and time order of rating.

3.0.1 Confidence Interval

Construct validity cannot usually be expressed as a single metric, but establishing an upper and lower bound can provide insight to the degree in which validity is obtained (Cronbach and Meehl, 1955). To that effect we have used the results of the "polychor" function, which provides an estimated covariance matrix for the maximum likelihood estimate of the latent bivariate Normal correlation, including the

Comparison	Sentiment	n	$\hat{\rho}$	Approx 95 CI
Scales 1 & 2	NP/AJ	1600	0.859	(0.841, 0.877)
Scales 1 & 3	NP/AJ	1600	0.886	(0.871, 0.901)
Scales 2 & 3	NP/AJ	1600	0.918	(0.910, 0.925)*

Table 3.1: Internal consistency of human raters

standard error of the estimate. Due to the large sample sizes of the data, the smallest comparison used is $n = 200$ and the largest being $n = 1600$, we chose to take advantage of the Central Limit Theorem and calculate confidence intervals using the Normal distribution method with the provided standard error.

In some cases, even though an estimation of the polychoric correlation was provided, there were errors that prevented the estimation of the covariance matrix. In those cases the confidence interval was estimated based on the estimate of ρ and the sample size using the CIr function for calculating the confidence interval for the Pearson's correlation from the "psychometric" package by Thomas D. Fletcher. These will be indicated by an asterisk.

3.1 Main Results

3.1.1 Internal Consistency of Raters Scores

Internal consistency of raters is a question that often is of concern, will raters perform similarly among tasks. Performing the tests of internal consistency allows us to check if the process of rating tweets is under control, and that raters are for the most part agreeing on the sentiment regardless of scale.

In Table 3.1, we pool information across all raters to see how they score the

same tweets according to each of the rating scales. The large polychoric correlations (> 0.8) indicate that the raters were performing consistently across tasks for scoring of the same tweet even though they were looking at different criteria for the scoring. The strongest polychoric correlation is between Score 2 and Score 3, with a $\hat{\rho}$ of 0.918. This should not be surprising as Score 3 begins with results from Score 2. The second highest relationship is between Score 1 and 3 with a $\hat{\rho}$ of 0.886, while the weakest is between Score 1 and Score 2, with a $\hat{\rho}$ of 0.859. Although all are different there can be seen to be high amounts of shared variability between the scoring types, providing evidence for consistency between ratings.

3.1.2 How Raters Scores Compare When Asked to Rate Simple vs. Complex Sentiments

Next we look at inter-rater reliability on our sentiment scales. The inter-rater reliability will provide and a baseline of how well raters agree on what sentiment is and of how well we can expect raters to compare to the sentiment analysis methods. These results are shown in Tables 3.2 and 3.3, below. For ease of discussion, we will consider not only $\hat{\rho}$, the polychoric correlation between measures, but also $100 * \hat{\rho}^2$, the percentage of shared variability between the rating methods.

3.1.2.1 Negative/Positive (NP) and Anger/Joy (AJ) Scales

Negativity and positivity are the most basic components sentiment can be broken into, and the lack of either would also indicate neutrality. As such almost all sentiment analyses look at these two. In this section, we will be examining how the raters scores compared to one another. Higher levels of agreement between raters would lend evidence for the content validity of sentiment analysis, that the sentiment in the text is something raters can agree on. From there we can compare the inter-rater correlations (Table 3.2) to the correlations between rater and computer sentiment

scores (Table 3.4). Similar levels of polychoric correlation between the raters and the sentiment analysis methods will indicate that both raters and systems capture a similar amount of an underlying construct, which supports the construct validity of sentiment analysis.

More nuanced than the negative and positive sentiment, anger and joy are subsets of these sentiments. In Table 3.3 we examine how the raters scores compared to one another. Higher levels of agreement between raters would lend evidence for the content validity of sentiment analysis for the sentiments anger and joy. Then we will compare the inter-rater scores to the rater-computer sentiment scores (Table 3.4). Again, similar levels of polychoric correlation between the raters and the sentiment analysis methods will indicate that both raters and systems agree on what a sentiment ought to be, which supports the construct validity of sentiment analysis for anger and joy.

Negative-Positive Scale 1			
	BC	BA	GM
BC	–		
BA	0.7573 (0.0017)	–	
GM	0.9118 (0.0004)	0.7502 (0.0018)	
TJ	0.8816 (0.0005)	0.7680 (0.0015)	0.9275 (0.0002)

Negative-Positive Scale 2			
	BC	BA	GM
BC	–		
BA	0.8643 (0.0006)	–	
GM	0.9081 (0.0003)	0.9423 (0.0001)	–
TJ	0.8812 (0.0004)	0.9225 (0.0003)	0.9349 (0.0002)

Negative-Positive Scale 3			
	BC	BA	GM
BC	–		
BA	0.8261 (0.0011)	–	
GM	0.8933 (0.0004)	0.9287 (0.0003)	–
TJ	0.8746 (0.0006)	0.9308 (0.0002)	0.9438 (0.0001)

Table 3.2: Inter-rater reliability for human raters on NP scales (Row and Column lables are raters' initials)

Anger-Joy Scale 1			
	RW	SM	LP
RW	–		
SM	0.8058 (0.0013)	–	
LP	0.7741 (0.0015)	0.8270 (0.0010)	–
KW	0.7527 (0.0023)	0.8217 (0.0015)	0.7543 (0.0022)

Anger-Joy Scale 2			
	RW	SM	LP
RW	–		
SM	0.8312 (0.0008)	–	
LP	0.7142 (0.0019)	0.7365 (0.0017)	–
KW	0.6292 (0.0037)	0.5904 (0.0041)	0.6839 (0.0030)

Anger-Joy Scale 3			
	RW	SM	LP
RW	–		
SM	0.7805 (0.0012)	–	
LP	0.6558 (0.0025)	0.6612 (0.0024)	–
KW	0.6833 (0.0029)	0.6083 (0.0037)	0.7253 (0.0024)

Table 3.3: Inter-rater reliability for human raters on AJ scales (Row and Column lables are raters' initials)

3.1.2.2 Inter-rater Reliability

Table 3.2 and 3.3 present results for inter-rater reliablity. Although no two people are the same, for a construct like sentiment to be valid, there should be shared ability to identify that construct. Seeing how well the raters can agree on a topic can shed light on the validity of sentiment analysis—that is, we can have credibility that we are fairly representing the sentiment of a text (Cronbach and Meehl, 1955). It can be seen that the negative and positive raters for the most part agree, with their lowest level of agreement being a polychoric correlation of 0.75 and their highest level being 0.94. If we consider this on the level of $\hat{\rho}^2$, this means that for all levels of scoring they have between 56.2% to 88.4% shared variability on what negative and positive sentiment is in a text. It should also be noted that the lowest amount of shared variability consistently came from rater BA on Scale 1. These results give a fair amount of confidence in the content validity of sentiment analysis for negative and positive sentiment.

On the other hand raters assigned to the anger and joy ratings seemed to be more varied in scoring. Their highest level of agreement was 0.83, and lowest 0.59, with an associated shared variability between 34.8% to 69.1%. Overall it appears that with anger and joy, the content for sentiment analysis is less well agreed upon. If raters do not have a consensus of what the sentiment entails in a text, it is harder to support the construct validity of using these emotions for sentiment analysis.

3.1.3 Rater Scores Compared Sentiment Analysis Tools

Are Sentiment Analysis Tools measuring the same as Human raters? If there is a high correlation that indicates there is agreement between the automated tools and people's scoring. If intra-rater reliability and sentiment analysis methods have comparable $\hat{\rho}$ values, that means that the sentiment analysis methods are providing

Comparison	Sentiment	n	$\hat{\rho}$	Approx 95 CI
Scale 1 & SentiStrength	NP	800	0.846	(0.816, 0.876)
Scale 2 & Bing	NP	800	0.888	(0.867, 0.909)
Scale 2 & NRC	NP/AJ	1600	0.789	(0.764, 0.814)
Scale 2 & NRC	NP	800	0.881	(0.859, 0.903)
Scale 2 & NRC	AJ	800	0.634	(0.580, 0.688)
Scale 2 & SentiWord	NP	800	0.788	(0.755, 0.821)
Scale 3 & AFINN	NP	800	0.872	(0.848, 0.897)

Table 3.4: Polychoric correlations of assessments by human raters and by comparable sentiment analysis methods.

results comparable to the sentiment assessments of a random human. If $\hat{\rho}$'s are systematically higher for intra-rater comparisons than for rater-computer comparisons, that will indicate that sentiment analysis is not matching up quite as well with what humans think about tweets. Table 3.4 shows information comparing raters' scores with SA methods pooling information across all raters whose scores are comparable to those SA methods.

Table 3.4 shows that the sentiment analysis tool Bing aligns the best with human raters, Scale 2 and Bing sharing 78.6% of variability in their sentiment assessments. The next closest analog to human ratings is the AFINN method, before dropping to SentiStrength. NRC and SentiWord perform similarly, and not quite as well as the other tools overall. Though it should be noted that the overall rater-computer correlation for NRC is comparing four sentiments: negativity, positivity, anger and joy; while the rest look at only negative and positive sentiment. The high levels of polychoric correlation for these methods lend support to the construct validity of sentiment analysis, in that the raters and the methods seem to be measuring roughly the same concept.

Considering rater-computer correlations for NRC broken out by sentiments con-

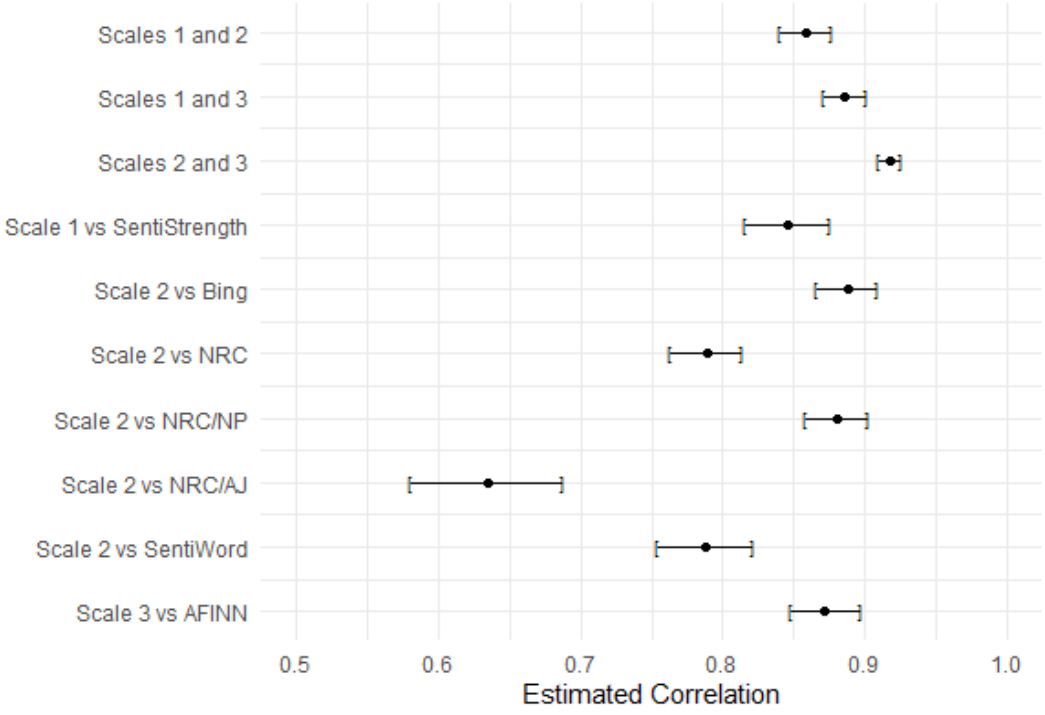

Figure 3.1: Estimates and confidence intervals for intra-rater correlations and rater-computer correlations.

sidered, there is a jump from 0.789 when considering all of the sentiments to 0.881 when only considering negative and positive sentiment. This indicates that the raters performed better at identifying negative and positive sentiment.

When considering both the intra-rater scores and the rater-computer sentiment analysis method scores, we do not notice any dramatic differences. Intra-rater scores for Scale 1 have shared variability between 56% to 89%, while the rater-computer comparison with SentiStrength has about 71.5% of shared variability. Intra-rater values for Scale 2 have between 74.7% to 88.7% of shared variability, while Bing shares 78.8% of variability with raters, NRC shares 77.6% with raters when considering NP scores, and SentiWord shares 62.1% of variability with raters. Finally intra-rater comparisons have between 68.2% to 89.1% of variability shared for Scale 3, while the AFINN method shares 76% of its variability with the raters.

SentiStrenth, Bing, NRC and AFINN look to have comparable levels of agreement with human raters in rating negative and positive sentiment. This provides evidence that these methods using negative and positive sentiment analysis have reasonable levels of construct validity. SentiWord looked to have lower levels of shared rater-computer variability than what was seen in the intra-rater's scores, and therefore seems to have less somewhat less validity as a measure of the NP sentiment construct.

The polychoric correlation between the raters and the sentiment analysis performed by NRC is the weakest seen, with only 40.2% of variability in scoring shared with the raters. In comparison the raters had between 37% to 60.9% of shared variability for Scale 2, the corresponding metric for the NRC method. The rater-computer sentiment analysis variability is comparable to the intra-rater variability, which indicates that the sentiment analysis is performing about as well as distinct humans at identifying the sentiments anger and joy. While performing similarly, it seems that both raters and NRC do not show the same level of content validity seen for negative/positive ratings. There is less clear agreement on what it means for a text to exhibit sentiments of joy or anger.

3.2 Secondary Results

The goal of these secondary analyses is to show that there's not any biases that could occur due to the treatment that raters were assigned to. The factors that could have contributed to a systematic difference are: if a rater received a training document with training intervention, the order of sentiments that raters were asked to evaluate, and the order in which tweets were presented to the raters.

3.2.1 How Priming Raters with Sentiment Related Words Impacts Scoring

Two sets of instructions were provided to the raters, either a neutral training document or a training document with descriptions of what would constitute a particular sentiment and specific examples of varying levels of that sentiment to "guide" or "prime" the rater (Weingarten et al., 2016). Looking at whether any bias was introduced in our results, whether intra-rater and rater-computer results depend on how the rater was trained, gives the opportunity to compare a more naturalistic rating to a rating that may be more similar to the systematic approach that is often used in computer-mediated sentiment analysis. Results for these analyses can be seen in Table 3.5.

			Guided		Unguided	
Comparison	Sentiment	n	$\hat{\rho}$	Approx 95 CI	$\hat{\rho}$	Approx 95 CI
Scales 1 & 2	NP	400	0.8245	(0.781, 0.868)	0.9298	(0.910, 0.950)
Scales 1 & 3	NP	400	0.8354	(0.793, 0.877)	0.9447	(0.933, 0.954)*
Scales 2 & 3	NP	400	0.9450	(0.933, 0.954)*	0.9387	(0.926, 0.950)*
Scale 1 & SentiStrength	NP	400	0.7775	(0.721, 0.834)	0.9049	(0.875, 0.935)
Scale 2 & Bing	NP	400	0.8706	(0.835, 0.906)	0.9103	(0.886, 0.935)
Scale 2 & NRC	NP	400	0.8687	(0.835, 0.903)	0.9108	(0.887, 0.935)
Scale 2 & SentiWord	NP	400	0.7733	(0.721, 0.826)	0.8137	(0.771, 0.857)
Scale 3 & AFINN	NP	400	0.8584	(0.821, 0.896)	0.9032	(0.875, 0.931)
Scales 1 & 2	AJ	400	0.745	(0.686, 0.803)	0.894	(0.860, 0.928)
Scales 1 & 3	AJ	400	0.786	(0.736, 0.836)	0.9397	(0.917, 0.963)
Scales 2 & 3	AJ	400	0.935	(0.922, 0.946)*	0.941	(0.911, 0.971)
Scale 2 & NRC	AJ	400	0.762	(0.707, 0.817)	0.546	(0.455, 0.636)

Table 3.5: Internal consistency of human raters and criterion validity with sentiment analysis measures—stratified by training.

3.2.1.1 Negative and Positive

The unguided polychoric correlation of Scale 1 and Scale 2, as well as Scale 1 and Scale 3, are stronger than the guided counterparts, as seen in Table 3.5. Meanwhile, the polychoric correlation between Scale 2 and Scale 3 are very similar for both groups.

Notice that the unguided polychoric correlations are consistently higher than the guided for the comparisons of the rater's score to the sentiment analysis tools as well. Coupled with the unguided internal scores also tending to be stronger than those for the guided raters, there exists the possibility that the guided raters are experiencing cognitive overload from the exposure to priming words.

Guided raters for negative and positive sentiment seem to consistently produce less stable results than unguided raters, both within their own scores and in their correlations with SA methods. In all, this suggests that raters already have an intrinsic sense for what negative and positive sentiment is, increasing our confidence in the validity of the NP sentiment construct. Moreover, the higher rater-computer $\hat{\rho}$'s for the unguided group suggests that the sentiment analysis methods surveyed are already good analogs to what naturalistic evaluation of sentiment in our sample of tweets would indicate.

3.2.1.2 Anger and Joy

Interestingly, the opposite trend exists when looking at the internal scores for the raters examining anger and joy. The guided raters have stronger polychoric correlation for Scales 1 and 2 as well as for Scales 1 and 3, with a similar statistic for Scales 2 and 3 for both guided and unguided raters. It was previously noticed that the intra-rater correlations for the anger and joy raters were not as high as those for

the negative and positive raters. This suggests that raters may not have as natural a sense for anger and joy ratings, which is consistent with previous results relating to the anger and joy content validity section.

Considering the rater-computer comparisons, again we see that the guided raters have a stronger polychoric correlation than the unguided raters when comparing to the NRC method of sentiment analysis. With the sentiments anger and joy, training seems to have a real positive effect on how much raters match with the sentiment analysis method. These ratings aren't necessarily capturing what people think of when they think of anger and joy, but training people on the rating task can make them match the computers better on this, if not on the other rating tasks.

To summarize, training with word priming in the negative and positive sentiments made raters less consistent with their internal scoring and with the sentiment analysis methods when compared to training without word priming. However, the reverse is observed in the anger and joy group, where the guided raters had better internal consistency and were better analogs to the sentiment analysis method. This suggests that our raters have a better sense of negative and positive sentiment than they do anger and joy. Our previous conjecture that there is a lack of content validity for assessments of anger and joy sentiments in a short text like a tweet are consistent with what we see here.

			Negative-First		Positive-First	
Comparison	Sentiment	n	$\hat{\rho}$	Approx 95 CI	$\hat{\rho}$	Approx 95 CI
Scales 1 & 2	NP	400	0.943	(0.926, 0.960)	0.832	(0.792, 0.873)
Scales 1 & 3	NP	400	0.948	(0.933, 0.964)	0.852	(0.815, 0.888)
Scales 2 & 3	NP	400	0.945	(0.934, 0.955)*	0.939	(0.926, 0.949)*
Scale 1 & SentiStrength	NP	400	0.887	(0.854, 0.920)	0.805	(0.753, 0.857)
Scale 2 & Bing	NP	400	0.886	(0.855, 0.916)	0.901	(0.874, 0.929)
Scale 2 & NRC	NP	400	0.873	(0.837, 0.909)	0.897	(0.870, 0.925)
Scale 2 & SentiWord	NP	400	0.784	(0.736, 0.832)	0.818	(0.778, 0.857)
Scale 3 & AFINN	NP	400	0.857	(0.820, 0.895)	0.892	(0.861, 0.923)

			Anger-First		Joy-First	
Comparison	Sentiment	n	$\hat{\rho}$	Approx 95 CI	$\hat{\rho}$	Approx 95 CI
Scales 1 & 2	AJ	400	0.903	(0.875, 0.930)	0.833	(0.786, 0.879)
Scales 1 & 3	AJ	400	0.941	(0.924, 0.958)	0.848	(0.805, 0.891)
Scales 2 & 3	AJ	400	0.938	(0.925, 0.948)*	0.947	(0.936, 0.957)*
Scale 2 & NRC	AJ	400	0.665	(0.594, 0.735)	0.627	(0.548, 0.706)

Table 3.6: Internal consistency of human raters and criterion validity with sentiment analysis measures—stratified by first sentiment rating task.

3.2.2 How Order of Sentiment Rating Impacts Scores

Next we assess the consistency of the raters based on which sentiment they were first asked to rate. Is there a condition that tends to more consistent results, or makes the learning process of rating easier? Results for these comparisons are given in Table 3.6.

3.2.2.1 Positive and Negative

The raters who scored a negative sentiment first seem to have somewhat stronger polychoric correlations for the Scale 1-2 and Scale 1-3 comparisons, and a similar score for the Scale 2-3 comparison (Table 3.6). This may suggest that those raters starting with negative sentiment were more stable in their overall ratings. A reason for this might be attributed something akin to negativity bias, where negative information tends to influence evaluations more than positive information (Ito et al., 1998). The fact that negativity makes a bigger impact on a person could imply that starting off with the negative rating set actually makes the learning process of identifying sentiment and rating it easier for the raters. Then, once having learned the process, those who started with negative sentiment more quickly learn to look for other sentiments and maintain their consistency for the whole rating process. (See, however, the next section on ordering which suggests that learning over time can't explain the differences seen here, making explanation of these differences more difficult.)

When comparing the rater-computer polychoric correlations based on order of first sentiment, the scores are fairly similar. This would be a good sign that there is negligible carry over from one rating to the next. The largest difference seen is between the Score 1 and SentiStrength. Here the negativity first group had a greater amount of shared variability than the positivity first group, 78.7% shared variability compared to 64.8%, which maybe worth further research.

In terms of internal consistency the negativity first group had the edge over the positivity first group. However, there was very little evidence of any differences with how well raters were matching sentiment analysis methods. That rater-computer polychoric correlations are roughly the same across conditions favors of the validity of the constructs being used in these computer sentiment analysis methods.

3.2.2.2 Anger and Joy

The raters who considered the sentiment anger first seem to have stronger polychoric correlation for the first two intra-rater comparisons and, again, a similar score for the final intra-rater comparison. This is the same pattern we saw when looking at the raters who were assigned to negativity and positivity, with the raters who first looked at negative or angry sentiment having better internal consistency for Scale 1-2 and Scale 1-3 comparisons than the raters who first looked at positive or joyful sentiment. This might support the idea that negative emotions are easier to identify than positive emotions (Ito et al., 1998), which could lead to the learning process of rating being easier for those who looked at negative sentiments first. Though again, see the lack of evidence for ordering effects in the next section.

The polychoric correlations comparing raters to the sentiment analysis method for those raters who considered anger first vs joy first are not much different from one another, and looking at confidence intervals further supports this lack of difference.

3.2.3 Raters Do Not Systematically Alter in Scoring over Time

In the previous section we looked to see whether the initial sentiment raters were asked to score affected the consistency of their sentiment scores. We found that peo-

Comparison	Slope Est.	Std. Err.	T-value	P-value
Scale 1 & SentiStrength	0.000051	0.00049	0.105	0.917
Scale 2 & Bing	-0.000053	0.00023	-0.236	0.813
Scale 2 & NRC	0.000032	0.00037	0.086	0.932
Scale 2 & SentiWord	0.000580	0.00045	1.267	0.206
Scale 3 & AFINN	0.000300	0.00026	1.190	0.235

Table 3.7: Regressions of discretized polychoric correlations against tweet ordering.

ple starting with negative emotions show marginally more consistency in their ratings from scale to scale, especially when Scale 1 is being considered. In this section, we assess if there seems to be any overall "learning curve", i.e. do raters' responses become more reliable over time. To check for a systematic change over time, small blocks of tweets for each rater will be used to calculate a polychoric correlation. These blocks of tweets are based on the order tweets were presented to each rater—so if a learning effect were to exist, we would expect to see raters' scores become more consistent over the course of their evaluations. Depending on the comparison there will be two hundred sets of either four or eight observations. From here a simple linear regression is performed using the polychoric correlation as the response variable and the observation order as the co-variate. A lack of a significant slope should be enough to say there is not a significant change in a rater's rating over time. Here, we limit our consideration to rater-computer correlations. Each rater was asked to rate according to all three scales, and ratings were done in order on the scales (i.e. raters evaluated all tweets relative to Scale 1, then relative to Scale 2 after finishing the Scale 1 evaluations, then relative to Scale 3 after finishing the Scale 2 evaluations).

As can be seen in Table 3.7, all rater-computer comparisons show slopes that are not distinguishable from zero, with quite large associated p-values. It appears as if raters never really get better or worse at the task. If some were learning the task faster than others, we'd expect to see some sort of learning effect in these regressions somewhere—lack of learning suggests no slope, which would attenuate learning

among some raters but not cancel it out. For the results to cancel across raters if some learning did happen, we would need other raters to get worse over time as well. Since there is no clear evidence of an order effect, we conclude that raters aren't getting meaningfully better at the task of evaluating tweet sentiment as they do more evaluations.

Chapter 4

Conclusions

4.1 Summary

Sentiment analysis is a tool that can be used to make sense of the growing textual data that surrounds us on a day to day basis, but needs to validated so that further uses can also be trusted. Using a time-stratified sample of data from Twitter, we collected sentiment scores on five computer-mediated methods of sentiment analysis, as well as on three human-scored scales created to mimic the computer-mediated methods under consideration. Raters' scores were compared internally and among the full pool of raters in order to evaluate the validity of the underlying sentiment constructs. This helps give an idea of whether the measures of sentiment are capturing features of the data that can be consistently identified.

Given that most sentiment analysis methods give ordinal ratings of a sentiment's strength, a more granular metric of validity is needed beyond the methods previously used to validate sentiment constructs and sentiment analysis methods. Polychoric correlation provides a method for construct validation of ordinally scored sentiment analysis. It produces clear and concise estimates of the inter-rater reliability and the similarity between rater's scores on multiple scales, as well as allowing comparison between human rater scores and scores on computer-mediated sentiment analysis

methods.

Our analysis found that there is good validity for methods that focus on negative and positive sentiments, with high intra-rater consistency, inter-rater reliability, and through comparisons between scores from human raters and those from computer-mediated methods. The sentiments anger and joy had lower levels of inter-rater consistency and the correlation between human scores and SA method scores was the weakest among the methods tested.

4.1.1 Key Results

Inter-rater reliability of the raters was quite high for negative and positive with shared variability between 56.6% to 88.4% for their scoring, supporting the content validity for these sentiments. This was not so for the anger and joy raters where shared variability was between 34.8% and 69.1%. The lower shared variability indicates that there is less consensus between raters as to what makes a text indicative of those sentiments.

When considering negative and positive sentiment, SentiStrenth, Bing, NRC and AFINN look to have comparable levels of agreement with human raters. This may indicate that there is a high construct validity to using the sentiment analysis tools that focus on negativity and positivity. SentiWord had lower levels of shared rater-computer variability than the previous methods.

Anger and joy sentiment analysis had a harder time keeping up with only 40.2% of variability in scoring shared between the raters and computer method scoring. This suggests there is not as much validity for these more nuanced sentiments.

4.1.2 Secondary Results

Scores for raters who were guided in identifying negative and positive words were found to have less internal consistency scores for raters who did not receive this

guidance. However, the reverse was observed in the anger and joy group, where the guided raters had better internal consistency and were better analogs to the sentiment analysis method.

This suggests that humans have a better sense of how to recognize negative and positive sentiment in text than they do for sentiments like anger and joy. We infer this because negative and positive sentiment analysis scores are more similar to scores by an unguided rater, and this pattern is fairly reliable across sentiment analysis methods. This pattern does not hold true for the anger and joy sentiments. For these sentiments, guided raters performed more like computer-mediated sentiment analysis methods than their unguided counter parts. This is consistent with previous conjecture that there is a lack of content validity for the expression of sentiments anger and joy in text.

The order of which sentiment was to be rated first for both negative/positive and anger/joy did not show much difference in correlations between raters' scores, internally or with computer-mediated methods. Additionally there did not seem to be a learning curve over the time raters performed—they did not become systematically better or worse at their task.

4.2 Limitations

Our study had a small sample group of raters, with only one rater per treatment combination. This creates a greater amount of variability than there would be if there were several raters in each treatment combination. This increased variability could limit the generalization of the results and restrict the reproducibility of these results. To correct for this, a replication of this experiment with a larger pool of raters should be done.

Additionally raters did not come from a random sample of the population. This presents a potential problem with independence of scores. Although it is difficult to truly randomly sample the population for volunteers, university research channels

can be used to obtain a sample of volunteers that are not associated with the proctor of the survey. Anecdotally, undergraduate psychology students tend to be in need of extra credit.

Another independence problem may come from raters giving three scores for each tweet at the same time. This may have over inflated the true internal relations between the scoring types. This could be rectified by instructing raters to only give one score at a time per text, then randomizing the order of a text and repeating the process. Furthermore, raters were asked to look at specified pairs of sentiment, either negative/positive or anger/joy. In a greater sample of subjects it would be prudent to look at a wider range of combinations of sentiment.

4.3 Further Directions

In the future we would like to design a more expansive experimental design and data collection process. First, a larger number of independent raters to provide more than a single observation per treatment category may confirm the results found in this study. In further research we also would not pair sentiments together and simply randomly assign two different sentiments to subjects. It would also be good to do a training task prior to data collection, to get raters used to the process. Finally, we would restructure the survey where raters only give one score for every text and with a randomized order so that raters don't fall into a habit of rating that could lead to a lack of independence between scores.

Concerning polychoric correlation, it would be worth spending time to investigate the distribution of this statistic more fully. With a formal distribution, more inferential statistics can be employed, instead of relying on a normality assumption that depends on the size of our data set.

Finally, a log-linear model approach may permit a wider range of inferential procedures than polychoric correlation. Further work will compare polychoric correlation and log-linear model approaches, to see what more can be learned through

this approach.

Appendix

Listed in this appendix are relevant technical details of this project.

A.1 R Code

In this section, code is included for study replication and further analysis.

A.1.1 Mining Twitter

Below is an example of the data used to mine posts from Twitter. A folder was specified to store the data and used as the working directory for the data collection process.

Twitter Developer limits downloads to 1500 every 15 minutes and will only download tweets closest to the time you ran the program downloading. Downloading 500 tweets every 5 minutes maximizes the number of tweets that can be downloaded at as many time points as possible using the tools available.

Check the size of the current .RData file and consider making new .RData files every 10-15 days to keep the data file from becoming to large and slowing down the download process. Between downloads, quickly rename the

```
tweets_df_2019.Rdata
```

to

```
tweets_df_2019_XX.Rdata
```

file. Still acting quickly create a new "empty"

```
tweets_df_2019.Rdata
```

using the commented out lines in the code. Doing this prevents the need to alter code.

A.1.1.1 Code

```
 # Load Required Packages
library(twitteR)
setwd("~/Tweets")
load("~/Tweets/tweets_df_2019.RData")
#load credentials
consumer_key <-   'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
consumer_secret<- 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
access_token <-   'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
access_secret <-  'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
setup_twitter_oauth(consumer_key ,consumer_secret,access_token ,access_secret)



#search string is "the" which is the most common used word in the English langue
search.string <- " " # testing if a literal " " will work as a search string
# "the, be, to, of, and" are the most common words I will use "the" "and" and "to"
no.of.tweets <- 500
```

```
tweets <- searchTwitter(search.string, n=no.of.tweets, lang="en")
tweets_2019new<- twListToDF(tweets)

tweets_df_2019<-rbind(tweets_df_2019, tweets_2019new)
#reset run line 42 and 43 AFTER changing file name of current tweets
# tweets_df_2019<-NULL
save(tweets_df_2019, file = "~/Tweets/tweets_df_2019.RData")
```

A.1.1.2 Automation of Mining

If using a Windows PC the program "Task Scheduler" can be set to automatically run the data collection code. The following steps are recommended for creating a scheduled task. This computer should be on and connected to the internet at all times.

1. Open Task Scheduler
2. Under Action select "Create Task." A new window will pop up.
3. Under the General Tab, name the task and give it a description.
   Select "Run whether user is logged on or not" and check "Run with highest privileges"
4. Under the Triggers Tab create a new trigger
   (a) Select a start time run the task daily starting at a time the user desires. This time can be in the past.
   (b) Select to run the task daily.
   (c) Select to repeat task every "5 minutes" for a duration of "Indefinitely."
   (d) Make sure that the Enabled box is checked off.
5. Under the Actions Table Create a new action.
   (a) Action should be "Start a Program."
   (b) Use the Browse to select the path to Rscript.exe (R program)
   (c) For "Add arguments", put the .R file that the downloading code was saved to.
   (d) For "Start in" give the path to the folder that the .R file with the downloading code was saved in.
6. Under Settings Tab

(a) Check, Allow Task to be run on demand.

(b) Check, Run task as soon as possible after a scheduled start is missed.

(c) Check, If the task fails, restart every 1 minute for 3 times.

(d) Check, Stop the task if it runs longer than 1 hour.

7. Save all changes

A.1.2 Sentiment Analysis

Below is the code to create the function to perform sentiment Analysis on strings text for the Bing, NRC, SentiWord and AFINN methods. Use the function “lapply” if there is more than one observation.

```
library(tidytext)
library(tidyverse)
library(glue)
library(stringr)
library(dplyr)
library(ggplot2)
library(wordcloud2)
library(stopwords)
library(sentimentr)

Mult_SA <- function(tweet_text) {

#load packages
library(tidytext)
library(tidyverse)
library(glue)
library(stringr)
library(dplyr)
library(ggplot2)
library(wordcloud2)
```

```
library(stopwords)
library(sentimentr)

#stopwords list
stopword <- stopwords(language = "en", source = "snowball")
more_word = rbind( "1", "2", "3","4","5","6","7","8", "9", "10")
stopword=data.frame(word = stopword)
more_word = data.frame(word= more_word)
stop_word <- rbind(more_word, stopword)

#load sentiword lexicon
load("~/2019/Thesis/Code/sentiwordlexicon.RData")

content<-NULL
content$str_text <- tweet_text
content$str_text = gsub("&amp", "", content$str_text)
content$str_text = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", content$str_text)
content$str_text = gsub("@\\w+", "", content$str_text)
content$str_text = gsub("[[:punct:]]", "", content$str_text)
content$str_text = gsub("[[:digit:]]", "", content$str_text)
```

```
content$str_text = gsub("http\\w+", "", content$str_text)
content$str_text = gsub("?", "", content$str_text)
content$str_text = gsub("\n", "", content$str_text)
content$str_text  = gsub("[\r\n]", "", content$str_text)

data_cont <-
  content$str_text %>%
  data_frame(text = content$str_text ) %>%
  unnest_tokens(word, text) %>%
  #   anti_join(stop_word, by=c("word"="word")) %>%
  #remove stopwords, comment if want to keep
  count(word)
pr_n <- sum(data_cont$n)

#######AFFIN
 data_sentAfinntemp <-
     content$str_text %>%
     data_frame(text = content$str_text ) %>%
     unnest_tokens(word, text) %>%
     inner_join(get_sentiments("afinn"))
```

```
  data_sentAnegtemp  <-data_sentAfinntemp %>%filter(score <= 0)
  data_sentApostemp  <-data_sentAfinntemp %>%filter(score > 0)

num_neg =      nrow(data_sentAnegtemp)
sum_neg =    sum( data_sentAnegtemp$score)
num_pos =      nrow(data_sentApostemp)
sum_pos =    sum( data_sentApostemp$score)

x <- data.frame("Num_neg_afinn" = num_neg, "Sum_neg_afinn" = sum_neg,"Num_pos_afinn"=num_pos,
                "Sum_pos_afinn"=sum_pos)

data_sentA <- x %>% mutate(average_neg_afinn = Sum_neg_afinn/Num_neg_afinn
                           ,average_pos_afinn = Sum_pos_afinn/Num_pos_afinn)
data_sentA <-as.tibble(data_sentA)
#####

##BING

  data_sentB <-
```

```
  content$str_text %>%
  data_frame(text = content$str_text ) %>%
  unnest_tokens(word, text) %>%
  inner_join(get_sentiments("bing")) %>%
  count(sentiment) %>%
  spread(sentiment, n, fill = 0)

  if (("negative" %in% colnames(data_sentB))==0){
   data_sentB<- mutate(data_sentB, negative = 0) }

  if (("positive" %in% colnames(data_sentB))==0){
    data_sentB <- mutate(data_sentB, positive = 0) }

if((nrow(data_sentB) ==0)){
  data_sentB <- add_row(data_sentB,negative=0,positive=0)}
  #if nothing is found then add a row

    data_sentB <-
      data_sentB %>%
      mutate(
```

```
    sentiment = positive - negative,
    emotion = positive + negative,
      ratio = negative/positive,
      mean_neg = negative/ pr_n,
      mean_pos = positive/ pr_n
        )

data_sentB <-
  data_sentB %>%
  dplyr::rename(
      Negative_Bing          = negative
    , Positive_Bing          = positive
    , Sentiment_Bing         = sentiment
    , Emotion_Bing           = emotion
    , Ratio_Bing             = ratio
    , Mean_Neg_Bing          =  mean_neg
    , Mean_Pos_Bing          = mean_pos)
######

######NRC
```

```
data_sentN <-
    content$str_text %>%
    data_frame(text = content$str_text ) %>%
    unnest_tokens(word, text) %>%
    inner_join(get_sentiments("nrc"))    %>%
    count(sentiment) %>%
    spread(sentiment, n, fill = 0)

  if (("anger" %in% colnames(data_sentN))==0){
      data_sentN <- mutate(data_sentN, anger = 0) }

  if (("anticipation" %in% colnames(data_sentN))==0){
      data_sentN <- mutate(data_sentN, anticipation = 0) }

  if (("disgust" %in% colnames(data_sentN))==0){
      data_sentN <- mutate(data_sentN, disgust = 0) }

  if (("fear" %in% colnames(data_sentN))==0){
      data_sentN <- mutate(data_sentN, fear = 0) }
```

```
if (("joy" %in% colnames(data_sentN))==0){
    data_sentN <- mutate(data_sentN, joy = 0) }

if (("negative" %in% colnames(data_sentN))==0){
  data_sentN<- mutate(data_sentN, negative = 0) }

if (("positive" %in% colnames(data_sentN))==0){
    data_sentN <- mutate(data_sentN, positive = 0) }

if (("sadness" %in% colnames(data_sentN))==0){
    data_sentN <- mutate(data_sentN, sadness = 0) }

if (("surprise" %in% colnames(data_sentN))==0){
    data_sentN <- mutate(data_sentN, surprise = 0) }

if (("trust" %in% colnames(data_sentN))==0){
    data_sentN <- mutate(data_sentN, trust = 0) }

if((nrow(data_sentN) ==0)){
```

```
   data_sentN <- add_row(data_sentN
                         ,joy=0,positive=0
                         ,anger=0, anticipation=0
                         ,disgust=0, fear =0
                         ,negative=0,sadness=0
                         ,surprise=0,trust=0)}
                         #if nothing is found then add a row

data_sentN <-   data_sentN %>%
 dplyr::rename(
   Joy_nrc                = joy
 , Positive_nrc               = positive
 , Anger_nrc               = anger
 , Anticipation_nrc               = anticipation
 , Disgust_nrc               = disgust
 , Fear_nrc               = fear
 , Negative_nrc               =  negative
 , Sadness_nrc               = sadness
 , Suprise_nrc               = surprise
 , Trust_nrc               = trust)
```

```
##########

######Senti word
  data_sentW <- content$str_text %>%
    data_frame(text = content$str_text ) %>%
    unnest_tokens(word, text) %>%  #spread out each word
    inner_join(sentiwordlexicon)  %>%
    count(sentiment) %>%
    spread(sentiment, n, fill = 0)

  if (("negative" %in% colnames(data_sentW))==0){
    data_sentW<- mutate(data_sentW, negative = 0) }

  if (("positive" %in% colnames(data_sentW))==0){
    data_sentW <- mutate(data_sentW, positive = 0) }

  if (("netural" %in% colnames(data_sentW))==0){
    data_sentW <- mutate(data_sentW, netural = 0) }
```

```
if((nrow(data_sentW) ==0)){
  data_sentW <- add_row(data_sentW,negative=0,positive=0,netural=0)}
  #if nothing is found then add a row

data_sentW$text<-content$str_text
data_sentW$n_word <-  pr_n

data_sentW <-
  data_sentW %>%
  dplyr::rename(
    Negative_SentiWord          = negative
  , Positive_SentiWord          = positive
  , Netural_SentiWord           = netural )

final_SA<-c(  data_sentA,data_sentB,data_sentN,data_sentW)
  return(final_SA)
}
```

A.2 Experimental Design

To validate the Sentiment Analysis tools, human raters are needed to simulate the scores given by the Sentiment Analysis for comparison. The way raters are to simulate the Sentiment Analysis ratings is described in Appendix A.3: Training Documents.

In this experimental design there are are three factors. The First factor is pairs of sentiment for the rater to consider, Negative and Positive or Anger and Joy. Negative and Positive sentiment are the most common sentiments used across sentiment analysis and also the most broad of categories. Anger and Joy similarly are two opposing sentiments that are of interest. The second factor is the type of training document given, guided or unguided training. In the guided training, raters are primed with sentiment words of a variety of strength to give them a sense of what high and low sentiment words may be. In the unguided training there is no priming. Finally and the order of of the sentiments to review, so that the order of which sentiment is rated first and second can be accounted for.

The treatments and are as follows:

Treatment	First Sentiment	Second Sentiment	Training
1	Negative	Positive	Guided
2	Positive	Negative	Guided
3	Negative	Positive	Unguided
4	Positive	Negative	Unguided
5	Anger	Joy	Guided
6	Joy	Anger	Guided
7	Anger	Joy	Unguided
8	Joy	Anger	Unguided

Two sets of data will be generated for the participants to review, one a stratified random sample of fifty tweets chosen from what the Sentiment Analysis tools classify as Negative Sentiment and fifty chosen from what the Sentiment Analysis

tools classify as Positive Sentiment. Stratification will be performed to ensure there is a variety of levels of sentiment in the sample, from each of the ten days randomly selected to be in the sample. Similarly a data set for Anger and Joy will be selected. The data set order will be randomized for each rating performed.

Eight subjects will each be randomly assigned to a treatment. They will read over the corresponding training document for the first sentiment and rate the tweets. Once finished with the first sentiment raters will be asked to take a fifteen minute break to help reduce carry over effect. If multiple raters are working at the same time they will be asked to refrain discussing their work until after the experiment. After their break the raters will be asked to read the second training document for the other sentiment that they are rating and rate a different randomized order of the tweets. When the second rating is complete the raters may ask the proctor questions.

A.3 Training Documents

Detailed below are examples of the training documents given to raters for scoring negative sentiment. Training documents for other sentiments are similar to these.

A.3.1 Unguided Training Document

A.3.1.1 Rating Sentiment of Text: Rater Section

When replicating the types of Sentiment Analysis used in this study, a human rater may use the following method to create comparison data for a desired string of text. You will consider one sentiment at a time, which will be selected by the proctor or by the system administering the survey. You will be asked to consider negative Sentiment.

Step 1

Before reading the any of the text, consider what would be negative sentiment, and reflect on what terms and phrases you might use to use to provide a list of examples of negative sentiment. Consider what strings of text might sound like with low amounts of negative sentiment, what might a string of text sound like with high levels of negative sentiment.

Note: Terms that are slang, misspellings, abbreviations, and acronyms should still be evaluated for negative sentiment.

Step 2

Read over the string of text.

Provide a score from 0 to 5 to describe how much negative sentiment you think is contained in the text.

A score of 0 indicates there is absolutely no negative sentiment in the text.

A score of 1 indicates there is a small amount of negative sentiment in the text.

A score of 3 indicates there is a moderate amount of negative sentiment in the text.

A score of 5 indicates that the text is extraordinarily laden with negative sentiment.

Step 3

Identify the terms in the text that you find to express negative sentiment and record the number of terms.

Step 4

Looking at the terms identified to contain negative sentiment rate each on a scale from 1 to 5.
A score of 1 indicates there is a small amount of negative sentiment in the term.
A score of 3 indicates there is a moderate amount of negative sentiment in the term.
A score of 5 indicates that the term is extraordinarily laden with negative sentiment.
Add up the scores for each term for one overall score.

Repeat steps 2-4 for all remaining texts.

Once finished assigning scores to all texts you may be asked to repeat the process with a different sentiment.

A.3.1.2 Clarifying Terminology

In the context of this survey, "term" indicates a word or distinct acronym with a space on either side.
In the context of this survey, "text" indicates the entire collection of terms for each observation. The text may be as short as a single term, or several sentences.

A.3.2 Guided Training Document

A.3.2.1 Rating Sentiment of Text: Rater Section

When replicating the types of Sentiment Analysis used in this study, a human rater may use the following method to create comparison data for a desired string of text. You will consider one sentiment at a time, which will be selected by the proctor or by the system administering the survey. You will be asked to consider negative Sentiment.

Step 1

Before reading the any of the text, consider what would be negative sentiment, and reflect on what terms and phrases you might use to use to provide a list of examples of negative sentiment. Consider what strings of text might sound like with low amounts of negative sentiment, what might a string of text sound like with high levels of negative sentiment. A term that may be considered to be a low level in negative sentiment might be the term "uncertain". A term may that may be considered to be high levels of negative sentiment may be "bitch".

Note: Terms that are slang, misspellings, abbreviations, and acronyms should still be evaluated for negative sentiment.

Step 2

Read over the string of text.

Provide a score from 0 to 5 to describe how much negative sentiment you think is contained in the text.

A score of 0 indicates there is absolutely no negative sentiment in the text.

A score of 1 indicates there is a small amount of negative sentiment in the text.

A score of 3 indicates there is a moderate amount of negative sentiment in the text.

A score of 5 indicates that the text is extraordinarily laden with negative sentiment.

Step 3

Identify the terms in the text that you find to express negative sentiment and record the number of terms.

Step 4

Looking at the terms identified to contain negative sentiment rate each on a scale from 1 to 5.

A score of 1 indicates there is a small amount of negative sentiment in the term. E.g. the term "sorry."

A score of 3 indicates there is a moderate amount of negative sentiment in the term. E.g. the term "obnoxious."

A score of 5 indicates that the term is extraordinarily laden with negative sentiment. E.g. the term "bastard."

Add up the scores for each term for one overall score.

Repeat steps 2-4 for all remaining texts.

Once finished assigning scores to all texts you may be asked to repeat the process with a different sentiment.

A.3.2.2 Clarifying Terminology

In the context of this survey, "term" indicates a word or distinct acronym with a space on either side.

In the context of this survey, "text" indicates the entire collection of terms for each observation. The text may be as short as a single term, or several sentences.

References

Michelle Annett and Grzegorz Kondrak. A comparison of sentiment analysis techniques: Polarizing movie blogs. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 25–35. Springer, 2008.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

Tawunrat Chalothorn and Jeremy Ellman. Sentiment analysis of web forums: Comparison between sentiwordnet and sentistrength. The 4th International Conference on Computer Technology and Development, 2012.

Xinyu Chen, Youngwoon Cho, and Suk Young Jang. Crime prediction using twitter sentiment and weather. In *2015 Systems and Information Engineering Design Symposium*, pages 63–68. IEEE, 2015.

Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

John Fox. Polycor: polychoric and polyserial correlations. *R package version 0.7-5*, 2007. URL `http://CRAN.R-project.org/package=polycor`.

P Haseena Rahmath and Tanvir Ahmad. Sentiment analysis techniques—a compar-

ative study. *International Journal of Computational Engineering & Management*, 17(4), 2014.

Tiffany A Ito, Jeff T Larsen, N Kyle Smith, and John T Cacioppo. Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, 75(4):887, 1998.

Norman Lloyd Johnson and Samuel Kotz. Distributions in statiscs; continuous multivariate distributions. Technical report, 1972.

Robbert Jongeling, Subhajit Datta, and Alexander Serebrenik. Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 531–535. IEEE, 2015.

Fajri Koto and Mirna Adriani. A comparative study on twitter sentiment analysis: Which features are good? In *International Conference on Applications of Natural Language to Information Systems*, pages 453–457. Springer, 2015.

Kalev Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 2011.

Bing Liu. Opinion mining and sentiment analysis. In *Web Data Mining*, pages 459–526. Springer, 2011.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis-a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, 2018.

Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

Ulf Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.

Steven Pinker. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Penguin, 2018.

Julia Silge and David Robinson. *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc., 2017.

Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, 2017.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

Evan Weingarten, Qijia Chen, Maxwell McAdams, Jessica Yi, Justin Hepler, and Dolores Albarracín. From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, 142(5): 472, 2016.

Gary E Weissman, Lyle H Ungar, Michael O Harhay, Katherine R Courtright, and Scott D Halpern. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics*, 89:114–121, 2019.

Kasper Welbers, Wouter Van Atteveldt, and Kenneth Benoit. Text analysis in r. *Communication Methods and Measures*, 11(4):245–265, 2017.