Scaling-Up Assessment from a Contextual Behavioral Science Perspective: Potential Uses of Technology for Analysis of Unstructured Text Data

Olga V. Berkout

Department of Psychology, Texas A&M-Corpus Christi

Angela J. Cathey

ENSO Group, Behavioral Science in the 21st Century

Karen Kate Kellum

Department of Psychology, University of Mississippi

Corresponding author: Olga V. Berkout

Present/permanent address: Texas A&M Corpus

6300 Ocean Drive,

Corpus Christi, TX 78412

Email: oberkout@gmail.com

Phone: (361) 825-5994

Fax:(361) 825-6098

ABSTRACT

With technological advancement, we have increased access to unstructured text data and the means to analyze it. Such information is available from electronic health records, blogs, social media posts, and other sources and is being used in business applications and social science research. These approaches can open up new areas for analysis, summarize massive amounts of information, and provide rapid feedback. Contextual behavioral science research in this area remains extremely limited. This manuscript provides an overview of techniques suitable for beginners interested in such analyses, potential applications, and resources for learning more.

Keywords: technology; contextual behavioral science; assessment; natural language processing

Scaling-Up Assessment from a Contextual Behavioral Science Perspective: Potential Uses of Technology for Analysis of Unstructured Text Data

Contextual behavioral science (CBS) concerns itself with the prediction and influence of behavior to support enriched meaningful living (Hayes, Barnes-Holmes, & Wilson, 2012). Along with the influence of environmental contingencies, CBS acknowledges that human behavior may be shaped by rules (Guinther & Dougher, 2015). Acceptance and Commitment Therapy (ACT), a widely used therapeutic approach from the CBS perspective, aims to reduce dysfunctional reliance on rules and strengthen contact with natural environmental contingencies (Biglan, 2009). This response to environmental contingencies, rather than difficult thoughts and feelings or rigid rule application, has been called psychological flexibility. ACT particularly emphasizes engagement in valued living, behaving in a manner that brings us closer to things that are meaningful and important (Wilson & Sandoz, 2008). Psychological constructs are viewed as useful to the extent that they support key goals of understanding, predicting, and modifying behavior (Hayes et al., 2012).

Psychological constructs considered important within CBS have most commonly been assessed using questionnaires. Questionnaires are inexpensive, easy to administer, and allow for comparison to others. Furthermore, self-report measures of acceptance, mindfulness, and psychological flexibility have been found to be useful predictors of key behaviors in psychological treatment (Cederberg, Cernvall, Dahl, von Essen, & Ljungman, 2015; Jordan, Wang, Donatoni, & Meier, 2014; Lloyd, Bond, & Flaxman, 2013; McCracken, 2013). However, reliance on questionnaires is not without potential issues. Important individual variability can be lost and respondents may struggle with accurate recall or be biased in their answers. Additionally, although questionnaire administration aims to allow for standardized comparison, this approach may miss important differences in experience and understanding. For example, Belzer and colleagues (2013) found that individuals with mindfulness training differed considerably from novices in their comprehension of items on a popular mindfulness inventory, suggesting that some measures may not be as useful without an adequate understanding of the covert behaviors being assessed (Atkins & Styles, 2016).

An alternative method is behavioral observation. Behavioral observation has a long history preceding modern CBS approaches and typically include coding and summarizing organism responses. Applied to verbal behavior, this involves use of trained coders to review audio, video, or written expressions and indicate when responses matching categories of interest show up. Many studies using this methodology have been conducted. Atkins and Styles (2016) examined how interview responses were related to psychological wellbeing and distress at a later point in time. Scholars focused on statements consistent with self-as-context (viewing experience as separate from oneself) and self-rules consistent with control or values. Control oriented self-rules were statements focused on avoiding negative consequences, whereas value oriented self rules focus on obtaining a desired experience. Atkins and Styles (2016) found that expressions of value oriented self-rules were predictive of greater wellbeing and less psychological distress at a later date and that self as context expressions were associated with lower subsequent symptoms of depression. Others have demonstrated that verbal expressions of acceptance and defusion during therapy sessions have been linked to subsequent reduction in tinnitus-related distress and impairment (Hesser, Westin, Hayes, & Andersson, 2009).

Similar approaches have been undertaken to capture individuals perceptions of their struggles and the therapeutic process. Bacon, Farhall, and Fossey (2013) conducted semi structured interviews to understand the experience of individuals with psychotic symptoms in ACT. This qualitative examination revealed a number of interesting distinctions in participants’ perceptions. Some viewed mindfulness as helpful, while others believed that it exacerbated symptoms. Scholars suggested that negative perceptions may be due to misperceptions of the purpose of mindfulness exercises as serving a relaxation function. Values based interventions were viewed more positively among this group. Generally, the overall intervention was seen as beneficial (Bacon et al., 2013). Analyses like these allow scholars to identify challenges within specific populations and work on developing ways to address them.

These approaches are less likely to lead to misunderstanding of participant perception compared to self-report. Participants are either observed or enabled to offer open responses to questions, allowing scholars to get a clearer picture of their perception of important constructs (as illustrated by Belzer et al., 2013). However, behavioral observation and qualitative data tend to require considerable resources. Coders must be trained in a standardized system and multiple coders may be needed to ensure classification agreement. If audio or video recordings are used, verbal expressions may need to be transcribed. This heavy resource load limits the amount of information to be examined. For example, Bacon and colleagues (2013) conducted their study on nine participants, a strategy scholars may undertake due to limited ability to gather and code this kind of rich verbal data. Sample size limitations may also make it more difficult to examine the impact of broad environmental changes (public policy, economic, green space access, etc.) on large numbers of individuals. When interested in such questions, scholars may turn to self-report, potentially missing important information that would have been uncovered in a less structured format. Additionally, even large scale surveys typically limit their assessment to sample sizes in the thousands and individuals are often assessed a limited number of times, losing temporal information.

At the same time our society is increasingly inundated with text data generated through social media, online forums, blog posts, and electronic health records. An estimated 68% of adults use Facebook, 28% Instagram, and 21% Twitter in the United States (Greenwood, Perrin, & Duggan, 2016). Individuals use these applications to share images, text, news stories, and articles with friends, family, or a broader base of followers. As many as six percent of Internet users report using reddit (an online discussion forum; Duggan & Smith, 2013). Some forums focus specifically on behavioral (e.g., the www.reddit.com mental health forum) and physical (e.g., the www.healingwell.com community forum) health difficulties. Electronic health record use has been rapidly increasing, although it is still less commonly used for behavioral compared to other health concerns (McGergor et al., 2015). In sum, there is a large and growing base of text data capturing individuals’ verbal expressions and provider perceptions. Furthermore, verbal interactions between therapists and clients may offer information on in session processes and serve as a useful addition to outcome questionnaires.

The purpose of this manuscript is to provide a brief overview of techniques suitable for behavioral scholars without backgrounds in computer programming. Each technique mentioned deserves a comprehensive review; however, our goal is to give scholars a sense of the options, rather than present all of the available information. The two programs mentioned, Python and R, are the most popular in big data analytics, with Python being more popular (Kaggle, 2017). Because it is more widespread, we will focus primarily on Python, although others programs may be mentioned when they provide a different functionality. The techniques we focus on were chosen using a combination of factors. First, we wanted to select techniques that were widely used and accepted within the field. Second, we wanted to limit ourselves to techniques and programs that would be suitable for beginners, as determined with input from our second author, who has considerable expertise in text analysis in business settings. Thus a number of widely used machine learning techniques, such as recurrent neural networks and random forests were not selected. Finally, the techniques discussed are meant to be illustrative of common approaches, rather than exhaustive, and our manuscript is not intended as a complete tutorial. We selected techniques that we hoped would encourage behavioral scientists to explore, but did not require an extensive understanding of computer science. Those looking to explore more complex applications may contact the second author, Angela Cathey, who has expertise in both computer science and CBS to serve as an intermediary on such projects.

**Working with Unstructured Text Data**

Although behavioral scholars are starting to analyze text data, these efforts within the CBS framework are extremely limited. However, techniques for analysis of unstructured text have received considerable attention in the fields of computer science and business analytics. Businesses use social media posts and online product reviews to assess and influence customer behavior (Bekmamedova & Shanks, 2014). Tools developed for these purposes may be adapted to be applicable to CBS.

**Programs to Consider**

As previously mentioned, R and Python are two of the most popular programs for big data analyses. Both are freely available. R is a software that was developed by and for statisticians (Tippmann, 2014). Python is a general purpose scripting language that has been applied to a range of problems including data analysis and machine learning (Ozgur, Colliau, Rogers, Hughes, & Myer-Tyson, 2017). Python tends to be more broadly applicable than R and is currently viewed as similar to R in its statistics capabilities (Brunner & Kim, 2016). Both R and Python are able to process text data and have large active user communities. Given its shallow learning curve and broader use (Bird, Klein, Loper, & Baldridge, 2008), we will steer our discussion towards Python; however, if scholars are more familiar and comfortable with R, they can implement similar tools within this program. Other programs and resources are described in our “Resources for Learning More” section. A comparison of the advantages and disadvantages of these and other programs mentioned is available in Table 5.

**Obtaining Data**

**Application Programming Interfaces.** The first step in undertaking an unstructured text research project is obtaining data. Scholars may already have the data they are interested in the form of interview transcripts or health records; however, they may also be interested in data available on social media, blogs, and online forums. This data may be accessed through application programming interfaces (APIs) or obtained through webscraping (Black, 2016). Social media websites commonly have API services, which may be used to extract data to facilitate the development of new applications that would interact with these platforms (Chen & Wojcik, 2016). An example illustrating obtaining data from reddit.com (a popular online forum) using Python is presented in Table 1. In our example the first thing we do is call the PRAW library (Boe, 2017), a specially developed toolkit allowing Python to get data from reddit using its API. We are asking PRAW to use our reddit account information to access reddit (as required by the forum) and requesting that the current top ten post titles from the Mental Health forum be returned. Readers will notice that Python uses a command line interface, in which we use text to ask the program to perform various tasks. An alternative that may be more familiar is a graphical user interface. Graphical user interfaces present images we click on to communicate with the program (an example is Microsoft Word). One alternative to obtaining data in this fashion is purchasing previously gathered social media data from resellers; however these can be prohibitively expensive (Mayr & Weller, 2017; McCay-Peet & Quan-Haase, 2017).

[Table 1 about here]

**Webscrapers.**Webscrapers are used to automatically browse the internet and extract text data, similarly to what one might do if looking for this information manually (Black, 2016; Glez-Peña, Lourenço, López-Fernández, Reboiro-Jato, & Fdez-Riverola, 2013). Python’s web scraper module, Beautiful Soup (Richardson, 2015), can be used to search for text, regular expressions, or other attributes in web documents and extract specific information (Nair, 2014). For individuals who would like a less programming heavy approach, Zapier offers an inexpensive alternative. Zapier allows scholars to extract information focusing on mention of a particular topic. It becomes active when a specified event (called a trigger) occurs on an online platform at which point it will perform an action specified by the user (Rahmati, Fernandes, Jung, & Prakash, 2017).

**Preprocessing data.** After text data have been acquired, it must be preprocessed to be suitable for subsequent analyses due to the limitations of computer processing. For example, we may ask that “crying” be categorized as a sad word and a program may not recognize “Crying” as having the same meaning. Scholars may remove capitalization from all words, so that this problem doesn’t occur (Denny & Spirling, 2018). For this same reason, scholars may also wish to remove punctuation (Denny & Spirling, 2018). Some programs will not recognize that “crying” and “crying!” should have a similar meaning. Scholars may tokenize words, labelling them as separate pieces of information, such that the sentence “I was crying” becomes “I”, “was”, “crying,” and the meaning of each word is considered separately (Madnani, 2007). Scholars also suggest that stop words (those which do not typically have meaning such as “the” and “a”) be removed to simplify analyses (Denny & Spirling, 2018). To further facilitate analyses, scholars may try to condense the information via stemming or lemmatizing to remove uninteresting variability. We may remove variability due to grammar, such as whether words are plural or gendered to more easily match these. Stemming does this by removing word endings, whereas lemmatizing compares words to a preexisting dictionary and ensures that real words remain (Effrosynidis, Symeonidis, & Arampatzis, 2017). Stemming appears to be more popular than lemmatization, potentially because it may be faster in large datasets (Chen, Thomans, & Hassan, 2015; Mahadik & Bharambe, 2015).

[Table 2 about here]

We give some illustrative simplified examples of preprocessing in Python in Table 2 (adapted from Caren, 2012 & Sentdex, 2015). We take a sentence (our post) in which we lower capitalized words, remove punctuation, remove stop words, and demonstrate stemming and lemmatization. Readers will notice that stemming shortens more words than lemmatization, sometimes in ways that are less sensical (attention becomes attent). To perform these techniques we are using Natural Language Toolkit (NLTK), a free Python library (Bird, Klein, & Loper, 2009). NLTK can perform a number of the language processing and analysis techniques and has the additional advantage of providing practice datasets and teaching resources due to being developed as a teaching tool (Bird et al., 2008). Although our example illustrates common approaches, preprocessing can vary depending on the dataset and intended analysis.

[Table 3 about here]

**Approaches**

**Sentiment analysis.** After data has been prepared, a number of techniques can be used to apply structure. A summary of these and their benefits and pitfalls is presented in Table 4. Sentiment analysis is a simple approach often used to get a sense of an individual’s view of something as positive or negative in the business world (Pawar, Shrishrimal, & Deshmukh, 2015). Practically, this is often applied to evaluate how new products or brands are perceived on social media or in reviews. There are two broad approaches used in sentiment analysis: machine learning and lexical dictionary based (Taboada, 2016). Lexical dictionaries are sets of words assigned a particular meaning. In the realm of sentiment analysis, lexical dictionaries might consist of positive words (like, awesome, great) and negative words (pointless, boring, worst; Hirschberg & Manning, 2015). Examining the occurrence of these words in text associated with a topic can give a sense of general attitudes towards it (Pawar et al., 2015). We offer a highly simplified example of sentiment analysis of our post in Table 3. We have two dictionaries representing positive (good, great, happy) and negative (sad, bad, cry) reactions and use this approach to tally how often the positive and negative terms occur in our post. Overall, the post describes an enjoyable experience with goat yoga and, consistently with this, includes two positive and zero negative terms from our dictionary. Machine learning approaches require more extensive understanding of computer programming and often use text examples labelled as positive or negative to train a classification algorithm (Taboada, 2016). For those interested in exploring this area, machine learning approaches can improve our ability to contextualize information and account for setting, changes in an individual's behavior, and particular characteristics of bodies of literature.

**Lexical dictionaries*.*** Although getting a sense of the extent to which something is viewed positively or negatively can be useful, this approach may be broadened to assess other ways of speaking. Lexical dictionaries can have larger numbers of terms consistent with various constructs of interest. Scholars may create their own dictionaries or draw on existing resources. WordNet Domains, a popular freely available dictionary has a WordNet Affect database, which provides words associated with specific emotional states and related psychological constructs (Bentivogli, Forner, Magnini, & Pianta, 2004; Strapparava & Valitutti, 2004). Dictionaries like these may be implemented in Python in a manner similar to sentiment analysis. We demonstrate with an example searching our ‘goat yoga’ post for a set of terms we have deemed consistent with mindfulness (noticing, present, attention; Table 3). We ask the program to count how many terms from the mindfulness dictionary occur in our post (in this case the word attention). This approach could be used to determine to what extent a text contains terms consistent with mindfulness.

Because of the popularity of lexical dictionary approaches a software allowing similar analyses outside of Python has been developed. Linguistic Inquiry Word Count (LIWC; Pennebaker, Booth, Boyd, & Francis, 2015) provides many dictionaries of words associated with particular psychological constructs, such as emotions, personality, and self focus (Tausczik & Pennebaker, 2009). LIWC has received psychometric support and has been used in a number of psychology studies (Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC has the advantage of being easier to implement than Python. However, paid dictionaries like this also keep their word lists private (otherwise anyone could replicate them without paying for the software). This creates a sort of black box, scholars might be uncertain of exactly what is happening. Such an approach may present challenges, particularly when word lists are applied to situations different from those in which they were developed.

**N-grams.** N-grams count occurrences of words in text without preserving their order (Schonlau & Guenther, 2016). Because of this the technique is also called “bag of words” (Chen & Wojcik, 2016). After preprocessing, scholars start by creating pairs (bigrams) or larger groups of words (trigrams for three words, and so on) to more adequately capture meaning (Schonlau & Guenther, 2016). We present an example of applying an n-gram approach to our ‘goat yoga’ post (Table 3). Similarly to the lexical dictionary approach, we have a list of words meant to represent mindfulness, except here these are word pairs: “paying attention”, “just noticing,” “being aware.” Our post is divided into word pairs and we compare these to the mindfulness ngrams and count the number of matches. In this case we match “paying attention” as a phrase from both mindfulness and our post. The advantage of n-grams over single words is that these may better capture what we are after. The term attention can be used in the phrase “stand at attention,” which is a military phrase not related to mindfulness. By specifying “paying attention” we are making it more likely that the phrase would refer to attentional focus to the present moment. This approach can also be used to assess for negation, as in the presence of “not happy” as indicating displeasure, whereas the search for a single word may just count the “happy” term as positive (Schonlau & Guenther, 2016).

**Regex.** Regular expressions (also called regex) are a format-based tool used to identify and extract a pattern in text (Schrenk, 2012). Through attending to pattern of use, rather than just occurrence, more pertinent information may be obtained. Dini and Bittar (2016) offer the example of the word “like” being relevant to emotion classification when it is used as a verb (“I like green eggs and ham”), but not as a preposition (“Green eggs are like ham”). Specifications of patterns of interest may be particularly helpful for social media analysis, which may include slang. We offer simple regex example in Table 3. Consistent with our theme, we are using Python and importing the re library for regular expressions (these may also be implemented in other languages). We are using a new post here, which describes feeling ill (“feeling sick”), while enjoying music (“sick beats”). Both of these phrases use the word sick, but if we are interested in physical illness, we would only be interested in the first one. To try to code for this we create two lists one of words that we think would represent illness and one that would be a specifier preceding a word from illness that we would expect to occur if a person was physically ill. Because we have the occurrence of feeling (expected for physical illness) and sick together, these are picked out by our regex code. At the same time, the phrase “sick beats” is not selected because sick occurs without the needed term.

**Artificial neural network.** CATPAC is a paid program providing an artificial neural network for analyzing text data, developed by Joesef Woelfel. CATPAC identifies key words in text and identifies similarity between words based on their use (Woelfel, 1998). This technique allows scholars to obtain clusters of commonly occuring words by frequency of co-occurrence (Klein, 2001). Clusters, called dendograms, typically focus on the most frequent words for simplicity, but may be modified to include terms of interest (Woelfel, 1998). In a sense, CATPAC allows scholars to conduct a sort of factor analysis, laying out word clusters which may be used to quickly create lexical dictionaries. As such, this tool may allow scholars to overcome the limitations inherent in the static nature of lexical dictionaries.

[Table 4 about here]

**Potential Applications in CBS**

Although these techniques generally not been used within the CBS framework (to our knowledge only a single CBS study by Collins and colleagues 2009 has been published), they have considerable potential. In fact, the relevance of natural language processing and other technologies which focus on ‘understanding’ human language was proposed by Greenway, Sandoz, & Perkins (2010). Sentiment analysis, which assesses the extent to which something is viewed as positive or negative (Pawar et al., 2015) could be used to evaluate response to a particular exercise or intervention. For example, Bacon and colleagues (2013) found that some individuals with psychotic symptoms appeared to view mindfulness exercises as less helpful because they misunderstood their function as relaxation. Negative responses to particular exercises may be explored further to determine contributing reasons. For example, negative response by African Americans to an exercise developed when working with Non-Hispanic Whites may suggest a need for adaptation. Sentiment analysis would not speak to the reason behind such responses, but may guide scholars to areas that need further attention.

Lexical dictionary and n-gram techniques may be used to assess language consistent with CBS constructs. Although limited research has been conducted in this area, some have found this approach to be useful. Collins and colleagues (2009) developed a mindfulness dictionary using input from an expert panel and analyzed substance use client responses on a participant feedback form. Linguistic Inquiry and Word Count (LIWC) was used to conduct analyses. Mindfulness language use was associated with greater use of present tense and body and affect related words, suggesting that participants who spoke in a manner more consistent with mindfulness appeared more focused on the present and aware of their emotional and physical states. As expected, scholars found negative associations between use of mindfulness language and alcohol use (Collins et al., 2009).

Regular expressions (regex) can be applied to further refine understanding of language and may be beneficial in analyzing psychotherapy interactions. Gallo and colleagues (2014) examined the potential for automated analysis of psychotherapy language to determine treatment fidelity. Scholars examined fidelity in a family intervention using an automated program focused on patterns in speech, similar to those that may be implemented using regex. Verbal expressions were coded by trained raters and the automated program for statements consistent with validation, open ended questions, encouragement to share anecdotes, and other language believed to reflect building a working alliance. Automated ratings were similar to those obtained by humans (correlation 0.84; Gallo et al., 2014). A similar process could be developed to offer feedback and support the development of CBS consistent skills. CATPAC, a program which can summarize text topics, may be used to summarize themes discussed in therapy. This approach may identify key concerns for specific populations, such as immigrants or refugees.

A notable benefit of automated analysis is that it requires considerably less manpower than manual coding. As such, scholars can analyze a massive amount of information and consider broad contextual variables. Public social media data can be used in public health and prevention efforts to identify the presence of experiential avoidance, distress, or valued living. These pieces of information can then be linked to location data and used to identify community factors related to values consistent psychologically flexible living. Such analyses may allow for examination of much broader contextual variables, such as policy decisions or events. Regression discontinuity paradigms could be used to evaluate the impact of policy changes, such as increases in the minimum wage. If validated programs are developed, automation could permit rapid feedback on data drawn from language use. Such analysis could inform CBS research and practice.

**Considerations and Concerns**

**Analyzing Text**

The approaches we describe in this manuscript allow scholars to make sense of the large amount of text data increasingly accessible in our society. Like any other analysis, this approach has strengths and weaknesses. Analysis of text data depends on decisions and input from those implementing this approach. When using lexical dictionaries, scholars may be generating language they believe relates to constructs of interest, as in Collins and colleagues (2009) expert created mindfulness dictionary.This is similar to the creation of a questionnaire, where a group of scholars may generate a number of items to represent a theoretical construct. As in the development of questionnaires, some of these terms may be more or less useful. Like any other approach, these should be validated to ensure that they are performing as intended. For example, expressions consistent with mindfulness in treatment sessions might be expected to correlate with self report measures of mindfulness and with associated constructs.

Analyses of text data may be anchored by other data sources. De Choudhury, Gamon, Counts, and Horvitz (2013) illustrate this in an examination of tweeting (posting on Twitter) in relation to depression. Scholars used Amazon’s Mechanical Turk (an online work platform) to identify individuals with depression based on scores on a validated depression screener and questions about diagnosis and treatment. Researchers identified individuals who had been experiencing symptoms for at least three months to a year prior to participating to have sufficient social media history. Participants provided access to their public Twitter profiles and the activity of those struggling with depression was compared to those who those without the difficulty. Scholars generated a dictionary of depressive symptom language based posts in an online depression forum for use in the study. This approach represents an alternative to expert input and was used to ensure that the analyses were capturing language as used within an online context. A number of group differences, such as greater late night tweeting and disclosure of symptoms, were demonstrated among those dealing with depression (De Choudhury et al., 2013). This study illustrates some of the ways that scholars might approach text analysis and strengthen methodology via integration of other validated measures.

Along with broader methodological considerations, CBS analyses have the additional consideration of serving the goals of prediction and influence of behavior. As such,

the utility of these analyses should be anchored to these goals. Should text data be predictive of a behavior of interest in a reliable fashion, it will be considered useful in the same manner as self-report measures of constructs, such as mindfulness and psychological flexibility. Inherent in this concept is the need to link data to outcomes outside this environment.

**Considering the Data Source**

Along with carefully considering their approach to the analysis, scholars should be mindful of the context within which the data were produced. Social media users may, for example, attempt to present themselves in a positive light (Bazarova & Choi, 2014). Not all users may be willing to be open about insecurities, negative emotions, and challenging life events. Users seeking information about stigmatizing conditions have been found to be more likely to do so through online searches, rather than social media posts (De Choudhury, Morris, & White, 2014). Health record data may vary depending on organizational practices and standards. Additionally, there are specific terms and means of communicating associated with various data sources. Social media is likely to have more slang and emojis than medical records. Medical records may contain diagnostic and billing codes, which would probably not be found in social media. Lexical dictionaries specific to these contexts have been developed and may be useful for those embarking on these analyses (Jensen, Jensen, & Brunak, 2012; Taboada, 2016).

Those interested in social media data should be familiar with limitations inherent in these platforms. API services may restrict the amount of information that can be retrieved and there are concerns that the data obtained through these may not be representative (Black, 2016; Ruths & Pfeffer, 2014). Ways in which information is extracted must also be carefully considered. A common practice is identifying posts on a topic through associated hashtags (use of a # followed by a term). However, users may select different hashtags for similar topics and some may not use these at all, potentially biasing the data obtained (Mayr & Weller, 2017). There are also notable demographic differences associated with different platforms. Reddit’s users are more likely to be young (18-29 years old) and male and Twitter users tend to have more education (Duggan & Smith, 2013; Greenwood, Perrin, & Duggan, 2016). Scholars should be aware of these as potential limitations to generalizability.

**Challenges for Automated Text Analysis**

Scholars who have interacted with automated telephone systems may have a sense of the challenges in automated understanding of language. Automated approaches have received growing support and continue to improve, but there are still cases in which they struggle in ways a human would not. If you saw the following social media post “Ugh! I was so happy this morning and now this email from my boss has me in the dumps! I feel like I’m bipolar,” you might be aware that the person is probably not referring to having received a Bipolar Disorder diagnosis. A simple automated program looking for “I’m bipolar” may not pick up on this idiosyncrasy. Sarcasm and irony continue to be difficult to classify (Taboada, 2016). Current use of automated text analysis approaches represents a tradeoff: scholars can analyze much more data, but its classification may struggle in ways a human coder would not.

**Ethical Concerns**

Scholars should also be mindful of the impact of their research on the users who produce the data and servers maintaining online websites. Researchers who webscrape online data must pause between requests. Failure to do so can prevent others from being able to access the site, similarly to the principle used in a denial of service cyber attack (Chen & Wojcik, 2016). Additionally, not all data that appears public may be considered as such. Facebook, for example, does not permit scraping of user profiles (Vitak, 2017).

There are also a number of ethical issues related to privacy. Recent news cases, such as concerns related to the use of social media data by Cambridge Analytica and the Facebook mood manipulation study highlight these issues. Briefly, the Facebook emotional contagion experiment involved an attempt to determine whether behavior on Facebook could be altered via the content that was shown to users (e.g., those who saw more negative posts would themselves post more negatively; Hunter & Evans, 2016). This study was conducted without the awareness or consent of users. Cambridge Analytica was a political firm assisting in President Trump’s 2016 campaign that also drew criticism for its use of Facebook data to target advertising due to issues regarding user consent (Granville, 2018). These cases emphasize a tension between the massive availability of information via such platforms and our awareness and input into the manner in which it is used.

Some have pointed out that research drawing on public forum and social media posts may not be using this information in the manner the original users intended (Hirschberg & Manning, 2015; Sloan & Quan-Haase, 2017). Social media terms of service agreements often allow for collection and analysis of data, although it may be argued that individuals are not truly understanding what they are agreeing to (Mittelstadt & Floridi, 2016). Some have also made the case that even though data may not appear identifiable, this does not necessarily mean this is the case (Metcalf & Crawford, 2016). Considerable disagreement continues to exist regarding practices using big datasets available online when Institutional Review Board members are surveyed (Vitak, Proferes, Shilton, & Ashktorab, 2017). Regulation of publicly available big data will likely continue to present challenges in research and industry contexts as regulatory agencies and social media platforms work to address mounting criticisms. For scholars working with data from individuals in the European Union, the General Data Protection Regulation (GDPR) attempts to address some of these issues. GDPR strengthens protections around user consent, right to withdraw consent, and transparency in data use (European Commission, n.d.)

**Resources for Learning More**

The tools described here are useful for beginners at this time; however, readers must keep in mind that this is a rapidly evolving field and new modules and programs are constantly being developed. As better options are made available, older ones may be relied upon less and become obsolete.

**Communities and Other Ways to Learn:**

Scholars may want to explore popular communities where issues associated with use of these programs, tutorials, and applications are discussed.

· Stack Overflow is a developer community where use of many languages, including Python and R, are discussed: https://stackoverflow.com

· Coursera is an online platform with courses covering any number of topics including text analysis, Python, and R: https://www.coursera.org

· KD Nuggets is a website/community focused on data science, including tutorials, resources, and applications. It does not emphasize text analysis specifically, but includes this approach along with others: https://www.kdnuggets.com

**Software Packages and Downloads Mentioned in the Paper**

**Open source programs.**

· Python may be downloaded at:[https://www.python.org](https://www.python.org/)

· R may be downloaded from: https://cran.r-project.org

**Libraries/modules.**

The easiest way to install these is via text commands in Python. However these links provide documentation regarding the use of these libraries

· PRAW Reddit API: <https://praw.readthedocs.io/en/latest/>

· Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/

· Regular expressions: <https://docs.python.org/2/library/re.html>

· Natural Language Toolkit (NLTK) [https://www.nltk.org](https://www.nltk.org/)and book http://www.nltk.org/book/

**Paid services.**

· Zapier may be accessed at: <https://zapier.com/apps/integrations>

· LIWC may be purchased from: https://liwc.wpengine.com/

· CATPAC may be purchased from http://www.galileoco.com/N\_catpac.asp

References

Atkins, P. W. B., & Styles, R. G. (2016). Measuring self and rules in what people say: Exploring whether self-discrimination predicts long-term wellbeing. *Journal of Contextual Behavioral Science*, *5*(2), 71–79. doi:10.1016/j.jcbs.2016.05.001

Bacon, T., Farhall, J., & Fossey, E. (2013). The active therapeutic processes of Acceptance and Commitment Therapy for persistent symptoms of psychosis: Clients’ perspectives. *Behavioural and Cognitive Psychotherapy*, *42*(04), 402–420. doi:10.1017/s1352465813000209

Bazarova, N. N., & Choi, Y. H. (2014). Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, *64*(4), 635–657. doi:10.1111/jcom.12106

Bekmamedova, N., & Shanks, G. (2014, January). *Social media analytics and business value: A theoretical framework and case study*. Presented at the 47th Hawaii International Conference on System Sciences, IEEE, Waikoloa, HI. doi:10.1109/hicss.2014.464

Belzer, F., Schmidt, S., Lucius-Hoene, G., Schneider, J., Orellana-Rios, C., & Sauer, S. (2013). Challenging the construct validity of mindfulness assessment—a cognitive interview study of the Freiburg Mindfulness Inventory. *Mindfulness*, *4*(1), 33–44. http://dx.doi.org/10.1007/s12671–012–0165–7.

Bentivogli, L. Forner, P., Magnini, B., & Pianta, E. (2004). Revising WordNet domains hierarchy: semantics, coverage, and balancing. In G. Serasset, S. Armstrong, C. Boitet, A. Popescu-Belis, & D. Tufis Eds. *Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources* (pp. 101-108). Geneva, Switzerland: HAL.

Biglan, A. (2009). Increasing psychological flexibility to influence cultural evolution. *Behavior and Social Issues*, *18*(1), 15-24. doi:10.5210/bsi.v18i1.2280

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O’Reilly Media, Inc.

Bird, S., Klein, E., Loper, E., & Baldridge, J. (2008, June). *Multidisciplinary instruction with the Natural Language Toolkit*. Paper presented at the Third Workshop on Issues in Teaching Computational Linguistics-TeachCL ’08. Columbus, Ohio, USA

Black, M. L. (2016). The World Wide Web as complex data set: Expanding the digital humanities into the twentieth century and beyond through internet research. *International Journal of Humanities and Arts Computing*, *10*(1), 95–109. doi:10.3366/ijhac.2016.0162

Boe, B. (2017). Python Reddit API Wrapper (PRAW). Retrieved from https://praw.readthedocs.io/en/latest/

Brunner, R. J. & Kim, E. J. (2016). Teaching data science. *Procedia Computer Science*, *80*, 1947–1956. doi:10.1016/j.procs.2016.05.513

Caren, N. (2012) An introduction to text analysis with Python. Retrieved from http://nealcaren.web.unc.edu/big-data/

Cederberg, J. T., Cernvall, M., Dahl, J., von Essen, L., & Ljungman, G. (2015). Acceptance as a mediator for change in Acceptance and Commitment Therapy for persons with chronic pain? *International Journal of Behavioral Medicine*, *23*(1), 21–29. doi:10.1007/s12529-015-9494-y

Chen, T.H., Thomas, S.W., & Hassan, A.E. (2015). A survey on the use of topic models when mining software repositories. *Empirical Software Engineering*, *21*(5), 1843–1919. doi:10.1007/s10664-015-9402-8

Chen, E. E. & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. doi:10.1037/met0000111

De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting Depression via Social Media. Seventh International AAAI Conference on Weblogs and Social Media; July 8-11, 2013; Massachusetts, USA. 2013.

De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. (2013, July). *Predicting depression via social media*. Paper presented at the Seventh International AAAI Conference on Weblogs and Social Media. Boston, MA.

De Choudhury, M., Morris, M. R., & White, R. W. (2014, May). *Seeking and sharing health information online*. Paper presented at the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI ’14. Toronto, Canada.

Collins, S. E., Chawla, N., Hsu, S. H., Grow, J., Otto, J. M., & Marlatt, G. A. (2009). Language-based measures of mindfulness: Initial validity and clinical utility. *Psychology of Addictive Behaviors*, *23*(4), 743–749. doi:10.1037/a0017579

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(02), 168–189. doi:10.1017/pan.2017.44

Dini, L. & Bittar, A. (2016). *Emotion analysis on Twitter: The hidden challenge*. Paper presented at the Tenth International Conference on Language Resources and Evaluation, Paris, France.

Duggan, M. & Smith, A. (2013). 6% of online adults are reddit users. Pew Research Center, Internet, Science and Tech. Retrieved from http://pewinternet.org/Reports/2013/reddit.aspx

Effrosynidis, D., Symeonidis, S., & Arampatzis, A. (2017). A comparison of pre-processing techniques for Twitter sentiment analysis. *Lecture Notes in Computer Science*, *10450*, 394–406. doi:10.1007/978-3-319-67008-9\_31

European Commission, (n.d.). Rules for the protection of personal data inside and outside the European Union. Retrieved from <https://ec.europa.eu/info/law/law-topic/data-protection_en>

Gallo, C., Pantin, H., Villamar, J., Prado, G., Tapia, M., Ogihara, M., … Brown, C. H. (2014). Blending qualitative and computational linguistics methods for fidelity assessment: Experience with the Familias Unidas preventive intervention. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 574–585. doi:10.1007/s10488-014-0538-4

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797. doi:10.1093/bib/bbt026

Granville, K. (2018, March 19). Facebook and Cambridge Analytica: What you need to know as the fallout widens. Retrieved from https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

Greenway, D.E., Sandoz, E.K., & Perkins, D.R. (2010, August). *Potential applications of relational frame theory to natural language systems*. Paper presented at 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. Yantai, China.

Greenwood, S., Perrin, A., Duggan, M. (2016). Social media update 2016 report. Pew Research Center, Internet, Science and Tech. Retrieved from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Guinther, P. M., & Dougher, M. J. (2015). The clinical relevance of stimulus equivalence and relational frame theory in influencing the behavior of verbally competent adults. *Current Opinion in Psychology*, *2*, 21–25. doi:10.1016/j.copsyc.2015.01.015

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual Behavioral Science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*,*1*(1-2), 1–16. doi:10.1016/j.jcbs.2012.09.004

Hesser, H., Westin, V., Hayes, S. C., & Andersson, G. (2009). Clients’ in-session acceptance and cognitive defusion behaviors in acceptance-based treatment of tinnitus distress. *Behaviour Research and Therapy*, *47*(6), 523–528. doi:10.1016/j.brat.2009.02.002

Hirshberg J. & Manning, C.D. (2015). Advances in natural language processing. *Science*, *349* (6245), 261-266. doi: 10.1126/science.aaa8685

Hunter, D., & Evans, N. (2016). Facebook emotional contagion experiment controversy. *Research Ethics*, *12*(1), 2–3. doi:10.1177/1747016115626341

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405. doi:10.1038/nrg3208

Jordan, C. H., Wang, W., Donatoni, L., & Meier, B. P. (2014). Mindful eating: Trait and state mindfulness predict healthier eating behavior. *Personality and Individual Differences*, *68*, 107–111. doi:10.1016/j.paid.2014.04.013

Kaggle (2017). The 2017 state of data science and machine learning. Retrieved from https://www.kaggle.com/surveys/2017.

Klein, H. (2001). Overview of text analysis software. *Bulletin of Sociological Methodology*, *70*, 53-66.

Madnani, N. (2007). Getting started on natural language processing with Python. *Crossroads*, *13*(4), 5–5. doi:10.1145/1315325.1315330

Mahadik, A. & Bharambe, A. (2015). Aspect based opinion mining and ranking: Survey. *International Journal of Current Engineering and Technology*, *5*(6), 3589-3592.

Mayr, P. & Weller, K. (2017). Think before you collect: Setting up a data collection approach for social media studies. In Sloan, L. & Quan-Haase, A. (Eds.), *the SAGE Handbook of Social Media Research Methods* (pp. 107-125). Thousand Oaks, CA: SAGE Publications Inc.

McCay-Peet, L. & Quan-Haase A. (2017). What is social media and what questions can social media research help us answer? In Sloan, L. & Quan-Haase, A. (Eds.), *the SAGE Handbook of Social Media Research Methods* (pp. 13-26). Thousand Oaks, CA: SAGE Publications Inc.

McCracken, L. M. (2013). Committed action: An application of the psychological flexibility model to activity patterns in chronic pain. *The Journal of Pain*, *14*(8), 828–835. doi:10.1016/j.jpain.2013.02.009

McGregor, B., Mack, D., Wrenn, G., Shim, R. S., Holden, K., & Satcher, D. (2015). Improving service coordination and reducing mental health disparities through adoption of electronic health records. *Psychiatric Services*, *66*(9), 985–987. doi:10.1176/appi.ps.201400095

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 1-14. doi:10.1177/2053951716650211

Mittelstadt, B.D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, *22*, 303-341. doi: 10.1007/s11948-015-9652-2

Nair, V.G. *Getting Started with Beautiful Soup*. (2014). Birmingham, U.K.: Packt Publishing.

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z. & Myer-Tyson, E. (2017). MatLab vs. Python vs. R. *Journal of Data Science*, *15*, 355-372.

Pawar, K.K, Shrishrimal, P.P., & Deshmukh, R.R. (2015). Twitter sentiment analysis: A review. *International Journal of Scientific & Engineering Research*, *6*(4), 957-964.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Rahmati, A., Fernandes, E., Jung, J. & Prakash, A. (2017) IFTTT vs. Zapier: A comparative study of trigger action programming frameworks. arXiv preprint Sep 2017 (arXiv:1709.02788)

Richardson, L. (2015) Beautiful Soup. Accessed at https://www.crummy.com/software/BeautifulSoup/

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. doi:10.1126/science.346.6213.1063

Schonlau, M., & Guenther, N. (2016). Text mining using n-grams. *SSRN Electronic Journal*. doi:10.2139/ssrn.2759033

Schrenk, M. (2012) *Webbots, Spiders, and Screen Scrapers, 2nd Edition*. No Starch Press, San Francisco, CA.

Sentdex (2015). NLTK with Python 3 for natural language processing. Retrieved from <https://www.youtube.com/watch?v=FLZvOKSCkxY&list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL>

Sloan, L. & Quan-Haase, A. (2017). A retrospective on state of the art social media research methods: Ethical decisions, big-small data rivalries, and the spectre of the 6Vs. In Sloan, L. & Quan-Haase, A. (Eds.), *the SAGE Handbook of Social Media Research Methods*(pp. 661-672). Thousand Oaks, CA: SAGE Publications Inc.

Strapparava, C. & Valitutti, A. (2004, May) *WordNet-Affect: an affective extension of WordNet*. Paper presented at theInternational Conference on Language Resources and Evaluation. Lisbon, Portugal

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, *2*(1), 325–347. doi:10.1146/annurev-linguistics-011415-040518

Tippmann, S. (2014). Programming tools: Adventures with R. *Nature*, *517*(7532), 109–110. doi:10.1038/517109a

Vitak, J. (2017). Facebook as a research tool in the social and computer sciences. In A. Quan-Haase & L. Sloan (Eds.), *The Sage Handbook of Social Media Research Methods.*(pp. 627-644). Thousand Oaks, CA: Sage.

Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics regulation in social computing research: Examining the role of Institutional Review Boards. *Journal of Empirical Research on Human Research Ethics*, *12*(5), 372–382. doi:10.1177/1556264617725200

Wilson, K.G. & Sandoz, E.K. (2008). Mindfulness, values, and the therapeutic relationship in Acceptance and Commitment Therapy. In S. Hick and T. Bein (Eds.). *Mindfulness and the Therapeutic Relationship* (pp. 89-106). New York, Guilford Press

Woelfel, J. (1998). *User’s guide: Captac II Version 2.0*. Rah Press.

Table 1. An Example of Obtaining Data Using an API

|  |  |  |
| --- | --- | --- |
| **Reddit API via Python using PRAW module** | | |
| This returns the names of the top ten mental health post titles for the reddit.com mental health subreddit | import praw  reddit = praw.Reddit (client\_id = ‘your\_client\_id’, client\_secret = ‘your\_client\_secret’, user\_agent=’your\_useragent’)  for submission in reddit.subreddit('mentalhealth').hot(limit=10):  ... print(submission.title) | Output:  Children and Mental Health  This hits close to home  Spent the evening depressed starting at the wall, but my floof was right there with me  🤷🏻‍♀️ 😔  Ok, I am determined to get \*something\* done today.  Can you feel lonely without feeling sad?  Is there a name for this form of deja vu type memories?  5 inspiring Ted talks about mental health; you must watch today  Hey everyone. Over the past couple of weeks I've been sharing some of my blog articles on here about mental health. If any of you have tried to look it up lately and the link isn't working, it's because my friends are posting on there now too, so I changed the name of the blog.  Need to vent |

Note: PRAW is a Python module for working with Reddit API (code adapted from Praw online documentation; Boe, 2017). Researcher’s personal information for reddit is not shared above, instead stand ins, such as “your\_client\_id” are used.

Table 2. Preprocessing Text Data

|  |  |  |
| --- | --- | --- |
| **Technique** | **Code** | **Output** |
| Lower capitalized words | import nltk  post="Took my first goat yoga class today! Great time, being in the moment and paying attention while playing with baby goats. Summer is good!"  post=post.lower ( )  print(post) | took my first goat yoga class today! great time, being in the moment and paying attention while playing with baby goats. summer is good! |
| Remove punctuation | import string  post=post.translate(None, string.punctuation)  print(post) | ‘took my first goat yoga class today great time being in the moment and paying attention while playing with baby goats summer is good’ |
| Tokenize | from nltk.tokenize import word\_tokenize  post=word\_tokenize(post)  print(post) | ['took', 'my', 'first', 'goat', 'yoga', 'class', 'today', 'great', 'time', 'being', 'in', 'the', 'moment', 'and', 'paying', 'attention', 'while', 'playing', 'with', 'baby', 'goats', 'summer', 'is', 'good'] |
| Remove stop words | from nltk.corpus import stopwords stop\_words=set(stopwords.words("english"))  post = [w for w in post if w not in stop\_words]  print (post) | ['took', 'first', 'goat', 'yoga', 'class', 'today', 'great', 'time', 'moment', 'paying', 'attention', 'playing', 'baby', 'goats', 'summer', 'good'] |
| Stem | from nltk.stem import PorterStemmer  ps = PorterStemmer()  post = [ps.stem(word) for word in post]  print(post) | ['took', 'first', 'goat', 'yoga', u'class', 'today', 'great', 'time', 'moment', u'pay', u'attent', u'play', u'babi', u'goat', 'summer', 'good'] |
| Lemmatize | from nltk.stem import WordNetLemmatizer  lem = WordNetLemmatizer()  post=[lem.lemmatize(w) for w in post]  print(post) | ['took', 'first', 'goat', 'yoga', 'class', 'today', 'great', 'time', 'moment', 'paying', 'attention', 'playing', 'baby', u'goat', 'summer', 'good'] |

Table 3. Simple Analysis Examples

|  |  |  |
| --- | --- | --- |
| **Technique** | **Code** | **Output** |
| Sentiment Analysis | negative\_words=0  positive\_words=0  post=['took', 'first', 'goat', 'yoga', 'class', 'today', 'great', 'time', 'moment', 'paying', 'attention', 'playing', 'baby', 'goats', 'summer', 'good']  positive=['good','great','happy']  negative=['sad','bad','cry']  for word in post:  ... if word in positive:  ... print word+' is positive'  ... positive\_words=positive\_words+1  .... if word in negative:  ... print word+' is negative'  ... negative\_words=negative\_words+1  print(positive\_words, negative\_words) | great is positive  good is positive  (2, 0) |
| Lexical Dictionary | mindfulness\_number = 0  mindfulness=['noticing','present','attention']  for word in post:  ... if word in mindfulness:  ... mindfulness\_number=mindfulness\_number+1  print(mindfulness\_number) | 1 |
| N-grams | from nltk import bigrams  mindfulness=[['paying','attention'],['just','noticing'],['being','aware']]  def getNGrams(post, n):  ... return [post[i:i+n] for i in range(len(post)-(n-1))]  postb=getNGrams(post, 2)  def mind\_count(list1, list2):  ... count=0  ... for x in list1:  ... for y in list2:  ... if x==y:  ... count=count+1  ... return count | postb=  [['took', 'first'], ['first', 'goat'], ['goat', 'yoga'], ['yoga', 'class'], ['class', 'today'], ['today', 'great'], ['great', 'time'], ['time', 'moment'], ['moment', 'paying'], ['paying', 'attention'], ['attention', 'playing'], ['playing', 'baby'], ['baby', 'goats'], ['goats', 'summer'], ['summer', 'good']]  mind\_count(mindfulness, postb) |
| Regex | import re  post='feeling sick today but also listening to sick beats from cardi b'  ill=['sick','ill','nauseous']  phys=['feeling','was','am']  ill2=('|'.join(map(re.escape, ill)))  phys2=('|'.join(map(re.escape, phys)))  regex\_str='('+phys2+')\s('+ill2+')'  p=re.compile((regex\_str), flags=re.I)  print(len(re.findall(p, post))) | print(len(re.findall(p, post)))  1 |

Note: Adated from Caren (2012) and Sentdex (2015).

Table 4. Comparison of Basic Techniques for Analyzing Text Data

|  |  |  |
| --- | --- | --- |
| **Approach** | **Benefits** | **Potential Pitfalls** |
| **Sentiment Analysis:**assesses the extent to which something is viewed as positive or negative | -simple to understand & implement  -vast array of broadly applicable and freely available dictionaries exist | -provides limited information  -words are examined without context  -not good at picking up sarcasm |
| **Lexical Dictionary:**assesses the extent to which document contains words consistent with particular constructs/categories | -similar to sentiment analysis but allows for more nuanced distinctions (for example words associated with particular emotions)  -many freely available dictionaries have been developed | -similar to sentiment analysis |
| **N-grams:** assesses the extent to which groups of words occur | -permit for presence of sets of multiple words to be examined  -may be more specific to the topic of interest | -similar to sentiment analysis |
| **Regex:** assesses for the occurrence of patterns in text | -allows for greater flexibility than other lexical approaches (can search for words near each other, words in order, etc.) | -can be cumbersome and more complicated to implement |
| **Machine Learning Algorithms** | -more accurate  -draw information from the available data | -require more extensive understanding of computer programming to implement  -require use of unfamiliar and complex algorithms |

Table 5. Comparison of Programs Used in Unstructured Text Analysis

|  |  |  |
| --- | --- | --- |
| **Program** | **Benefits/Uses** | **Potential Negatives** |
| **Python** | -open source  -most widely used program among data professionals  -most broadly applicable  -facilitates interdisciplinary collaboration  -active user community (provides tutorials, information, the ability to ask questions) | -may be somewhat less specialized for general statistics than R  -command line interface  -may require that lexical dictionaries be developed by the researcher |
| **R** | -open source  -second most popular behind Python  -active user community and associated benefits | -limited primarily to statistical applications  -similar to Python |
| **LIWC** | -well validated and well established  -easy to use  -large number of dictionaries already developed and validated for psychological constructs | -proprietary dictionaries: unclear how constructs are assessed  -paid dictionary, potentially limiting its accessibility |
| **CATPAC** | -allows one to use advanced technique for summarizing information in text  -less reliance on potentially erroneous human input | -paid program, may limit accessibility  -less flexibility compared to learning programming and machine learning algorithms |