SLANG (Summary Linguistic Analysis Guide) and its Applications to Military Text-Based Analysis

Julia Haines
4225 Logistics Ave, #S205
Wright-Patterson Air Force Base, OH, 45433
UNITED STATES OF AMERICA (USA)

Julia.Haines.1@us.af.mil

Key words: Big Data/Advanced Analytics, Linguistic Analysis, Artificial Intelligence, Sentiment Analysis, Topic Modeling, Opinion Mining

ABSTRACT

This presentation will be a demonstration of and discussion about the SLANG (Summary Linguistic ANalysis Guide) tool, developed as a text analysis application specifically for the Department of Defense construct. The Air Force possesses massive amounts of text-based data, with no consistent approach for translating it into quantitative results. Surveys, contracts, and many other data sources provide valuable information for decision makers, but it is resource-intensive to manually read through thousands of documents to generate information for decisional support. SLANG is an organically-developed, no-cost tool that automates textual analysis, modernizing the approach, so decision makers have access to information in a condensed, reliable format. It allows analysts to efficiently and effectively work with text-based data, with methods backed by thorough research on the latest techniques. This tool was created specifically for the military context, and addresses multiple data types, such as survey results, document reviews, literature in base newspapers or magazines, press releases and web-based commentary. SLANG provides one-click sentiment analysis, topic modeling, and linguistic charts, all based on best-practice methodologies. SLANG gives the analyst the ability to make informed decisions about sentiment dictionaries, ensuring a higher level of accuracy and reliability. It has been used for a variety of applications across the Air Force and larger Department of Defense since its development less than a year ago. SLANG is meant to be a living tool, so future versions may incorporate additional capabilities.

1.0 INTRODUCTION

"Once upon a time, surveys were a staple for every leader to solicit feedback and every company to assess engagement. But now, surveys are starting to look like diesel trucks collecting dust in the age of electric cars." This quote comes from an article in Harvard Business Review titled "Employee Surveys are still one of the best ways to Measure Engagement," published in 2018 (Judd, 2018). The article is written by Scott Judd, head of People Analytics at Facebook, and Eric O'Rourke, People Growth & Survey Analytics lead at Facebook, a large company often noted for cultivating workforce climate in the environment of Silicon Valley. Despite the recent emphasis by business leaders on cultivating company climate and sustaining employee wellbeing, the article argues as its headline states, that employee surveys are an invaluable source of information.

Tracking employee attitudes across a workforce is not only worthwhile for cultivating climate, but also for predicting retention, maximizing profit margins and identifying markers of success for the company as a whole (Huselid, 1995). While some companies, in recent years, have opted to forgo the survey distribution, taking these metrics passively instead through tracking internet usage, email response and social networks, many studies have shown that the mere act of distributing a survey and asking employees for feedback has direct positive impacts on workforce health and unity (Judd, 2018). Additionally, when employees perceive

that nothing is being done with or about the results of a survey, there tends to be a more negative impact than if the survey had never been distributed at all (Council, 2020).

As a result, there is a lot of interest in how to efficiently and effectively analyze and act upon results of employee engagement surveys. Kenny argues in *Harvard Business Review* that "Managers, especially those in large organizations, spend an inordinate amount of time and money measuring the satisfaction levels of their staff. Sections of HR are dedicated to running employee satisfaction surveys and making sure managers conduct frequent check-ins with their direct reports." (Kenny, 2020)

When surveys are distributed across the Air Force (AF), whether it be an employee engagement survey, a climate survey, or similar, significant resources are put towards the development, distribution and analysis of the survey. However, when open-ended questions are included on these surveys, respondent comments are generally under-utilized, more often treated as a source for pull-quotes rather than a data source in and of themselves. This is due to a lack of transparency and confidence in the accuracy of machine-aided methods such as sentiment analysis and topic modeling. This confidence reduces further when the text has special context, such as within the Air Force context. No model or methodology has been universally identified as ideal for this use case, nor has any model been universally adapted. The inconsistencies in approaches across analytical teams tasked with assessing the results of these surveys leaves data on the field.

Even though open-ended questions are often optional, response rates indicate that there is usually enough information to analyze without fear that a select few comments will carry the weight of an entire conclusion. For example, Facebook analysts recently wrote in Harvard Business Review that "when [Facebook] send[s] out a survey, we get a surprising volume of write-in comments: on average, 61% of our people submit their own feedback and suggestions, and each person touches on five distinct topics. It is clear that people take the survey seriously and want to be heard." (Judd, 2018) So, an objective of this research was to ensure that the data from those open-ended questions is analytically "heard."

The research behind the development of the SLANG tool quantifies the accuracy of some common sentiment analysis and topic modeling methods in order to gain a better understanding of the scope to which they should and can be applied. In order to investigate this question, various sentiment analysis packages and lexicons were implemented in R and applied to textual data from various surveys distributed across the Air Force. Accuracy was assessed via comparison with manual sentiment classifications noted by an expert team of reviewers familiar with the Air Force context. The results indicate that sentiment analysis methods alone are not sufficient when applied to this context, although various adjustments were also investigated to significantly improve accuracy. This implies that Air Force analysts tasked with analyzing textual survey data should be hesitant to apply fully automated sentiment analysis or topic modeling as the sole method for generating conclusions about the body of text as a whole.

SLANG can be applied to text-based data outside of workforce surveys and opinion solicitations. However, given that analysis of survey data was the primary drive behind the development of the tool, and has since been the primary application, that is the focus of this research and of the associated tool demonstration.

2.0 ANALYTICAL METHODS

2.1 Analytical Purpose

There is no clear "best" method for topic modeling or for sentiment analysis, since all models perform better and worse in different contexts, depending on the method itself and the data on which it was trained (Ribeiro et al., 2016). For example, a model trained on restaurant reviews may take a comment containing the word "fresh" and learn to group it with like-comments, assumedly positive reviews about a salad or fish-based dish using similar words like "crisp", "green", and "organic". However, if that same model is then expected

to classify movie reviews, it would be incorrect to assume that comments containing the word "fresh" and "crisp" are of the same topic, since in that context, "fresh" likely refers to the plot or cast, and "crisp" the production film quality. Therefore, even though "fresh" and "crisp" have positive connotations in both contexts, the model will perform much better in the realm of restaurant reviews than movies reviews. This is important to keep in mind not only across topic categories, but also within topic realms wherein one category of commentary may span different languages, different dialects, and different time frames in which jargon has shifted (Jagtap & Pawar, 2013).

For this reason, analysts applying topic modeling and/or sentiment analysis methods to data must be aware of the original use-case scenario, and caveat those assumptions which may not translate well to new data. Sometimes, it is possible to adjust an existing model or method for a new use-case by re-interpreting (manually or automatically) a few select words that are highly context-specific (Tan & Zhang, 2008). One can also vary the application of multiple methods to achieve aspect-level sentiment analysis. However, in order to quantify if and by how much an adjustment improves the "fit" of the model, one must obtain an awareness of the performance of the model before and after the adjustment by quantifying the accuracy or notionally checking a few varied text pieces. In order to holistically measure the performance of a model, a substantial portion of the data needs to have been manually sorted or classified into the different sentiments and/or topics. Because this is not always feasible, however, analysts may sometimes take a varied subset of the data, and if the results of the model seem to match or mirror notional logic, the application of the model, while not perfect, is deemed good enough.

Due to the widespread variety in the way that people talk, write, and even express sarcasm, and the room for different interpretations of a text even between two individuals, no model in linguistics will perform as well as models may in more predictable fields. Thus, "good enough" is relative. While far from perfect, the application of these models to long-form survey data may enable leaders to generate conclusions from employee feedback which are otherwise ignored, maximizing the return on investment in the survey and juicing the data for all it has to say. In the following experiment, advanced methods have been carefully applied and assessed for accuracy in an attempt to understand realistic accuracy expectations and identify shortfalls as potential areas of improvement. The result is a clearer understanding of which models and methods are capable of providing actionable, quantifiable and reasonably accurate conclusions drawn from previously muted qualitative linguistic data, specifically in the Air Force context.

When surveys are distributed across and within the Air Force, whether it be an employee engagement survey or a climate survey or the like, many resources are put towards the development, distribution and analysis of the survey throughout its lifecycle. This is evidenced by the existence of the Air Force Survey Office. However, when open-ended questions are included on these surveys, respondent comments are under-utilized, more often treated as a source for pull-quotes rather than a data source in of themselves (Jipa, 2019).

There are a number of pre-processing steps that are relatively constant across all linguistic analysis methodologies and applications. This includes steps such as stopword removal, stemming, Part of Speech (PoS) tagging, normalization, and tokenization (Clark, 2018). These are basic principles in linguistic analysis and can be researched separately if the reader is unfamiliar.

2.2 Sentiment Analysis

One way to adapt more complex analysis methods for survey analysis is to apply sentiment analysis, i.e. quantifying the degree to which a sentence or paragraph expresses positive, negative, neutral and/or other emotions/sentiments. This can be conceptualized as a classification and/or clustering method similar to the topic modeling described above, but instead of grouping aspects by topic association, they are instead grouped by sentiment association, whether it be positive and negative or, as found in the National Research Council Canada (NRC) Emotion Lexicon, association with 10 core emotions such as trust and anger (A. & Sonawane, 2016).

This can be done at multiple levels, depending on if the analyst is interested in whether the paragraph as a whole trends positive, whether a sentence trends positive, and even whether mentions of a topic within the sentence trends positive. The scale of the score differs by method, but for example, scores of +1 indicate positive sentiment, scores of -1 indicate negative sentiment, and scores near or at 0 indicate neutral sentiment. When these scores are computed at the topic level, it is called *aspect-level sentiment analysis*, a combination of topic modeling and sentiment analysis applications (Luo et al., 2016).

There are both supervised and unsupervised approaches to sentiment analysis. Support Vector Machines (SVM) and N-gram algorithms are used together for emotion identification of twitter messages (Almatarneh & Gamallo, 2019). Rules-based algorithms are generally easy to follow and easy to implement, but they are difficult to maintain since the rules need to be updated with a degree of consistency and the analyst needs to have a very active role in defining the rules being used. The other type of sentiment analysis algorithm, automatic algorithms, take longer to set up and train since they are based in machine learning, but are often more accurate and holistic in their results (Almatarneh & Gamallo, 2019). In order to first use the data to train the algorithm, it is fed through an n-grams or bag of words type process so that the machine can identify factors of the string which may contribute to the string's sentiment score (Almatarneh & Gamallo, 2019).

Because there are a wide variety of applications of sentiment analysis, there are also many different methods that are available to use, depending on the dataset and use case. There have also been studies assessing the differences between these methods, which have proven that there is no one "best" application. "The benchmark analyses reveal that there is no superior sentiment analysis method because all tools perform differently depending on the specific context they are applied on or depending on the corresponding data source on which they were trained." (Feine et al., 2019) Thus, an ideal sentiment analysis method must be selected not only based on the data with which it was trained, or based on, but also based on the perceived or calculated accuracy of the method when applied to specific data.

A study titled "SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods" published by EPJ Data Science Journal in 2016 found that two of the "best" methods for measuring numerical polarity in sentiment analysis, identifying positive, neutral, and negative comments, were VADER and AFINN. Both use a set of rules and heuristics to assess the degree to which a piece of text aligns with a given lexicon, and those lexicons are trained on social media data (Ribeiro et al., 2016).

Different, machine-learning based approaches developed by technology companies such as IBM, Microsoft and Google have been shown to perform better on varied datasets. With these methods, machine learning classification algorithms are used to predict the sentiment score of a piece of text. Thus, for this research question, the survey results can be applied to both rule-based and machine learning-based methods through open-source APIs (Corredera et al., 2017) and the webservice ifeel 2.0 (Araujo et al., 2016), as they did in the study of Chatbot Customer Service Sentiment Analysis (Feine et al., 2019).

In order to properly compare these methods, this research will standardize the sentiment scores obtained by each of these methods, and conduct correlation tests between the sentiment scores of different methods with those computed manually. This will reveal which, if any, of the sentiment analysis methods are valid for application to this type of data.

2.3 Topic Modeling

The SLANG tool allows user-defined topic modeling, an approach largely accepted when topics have lots of overlap in key ideas, such as is found in workplace satisfaction surveys. The analyst may want to distinguish that they want employee comments mentioning "telework" to be considered a climate topic, a communication topic, or a topic of its own to be paired with comments mentioning "work from home." However, automated techniques for sentiment analysis are effective when topics do not risk overlap, such as

Latent Dirichlet Allocation (LDA). Therefore, that approach has been researched for the military context and is detailed below, but is not currently implemented in the SLANG tool.

In this technique, comments are accumulated and treated as a bag of words, each with different probabilities, and then topics are derived that compose each comment. The assumption is that topics and words each have distributions underlying the text, and one can use those distributions to identify topics and the words associated with them. LDA is the most common approach, but requires the text to be transformed into a document term matrix and cleaned for punctuation, etc, before being applied (Koch, 2020).

In this research, LDA was applied with the package Gensim in Python, as well as with R. Essentially, the analyst must decide to what degree she wants the topic modeling to apply, how many words should be associated with each topic. This restricts the algorithm from identifying all text in the corpus as under one topic, or from splitting it into as many topics as there are comments. The ideal number for this choice depends on the size of the text and the questions being asked by the analyst. Next, the analyst may choose a topic mixture, i.e. what they expect the degree of topics to be distributed among those identified. Then, words start mapping to the topics and the model starts to learn (Chen, 2011; Clark, 2018; Wang, 2017).

3.0 RESULTS

3.1 Assumptions/Limitations

As with all linguistic analysis methodologies and applications, there are substantial assumptions which should be stated prior to definitive conclusions being drawn. Below, those assumptions have been generalized, but more specific assumptions relating to certain methodologies or mathematical principles are detailed at greater length in “Methodology”.

Sentiment analysis techniques are notorious for their inability to accurately and consistently detect sarcasm (Salas-Zárate et al., 2017). For example, take a respondent who writes “I think leadership is doing a great job, I really love how considerate they are of the team’s time when they’re drinking coffee half the time and delegating all their work the other half.” A sentiment analysis algorithm will see the words “love” and “considerate”, positive connotations, in association with “leadership”, and will likely classify this comment as positive.

Many are able to negate those words if preceded by “don’t” love or “not” considerate, but without those negators, and without any words that are negatively connotated, this comment will be incorrectly classified (Wang, 2017). However, on the whole, it is safe to assume that the majority of comments will not be sarcastic, and insights may be drawn about a comment database as a whole as long as analysts are aware of these limitations. For example, this example above will still teach a model to associate “leadership” with “coffee” and “delegate”, which are valuable insights to be aware of.

There is also research which shows that, in survey data specifically, those who bother to comment and write answers to free-response questions are usually a bit more disgruntled, while those satisfied with their workplace environment may leave those optional questions blank (Luo et al., 2016). However, there is also research which shows otherwise (Jipa, 2019). Either way, this is something to keep in mind, that any results drawn from linguistic analysis of survey data come only from the population of respondents who had something to say, not from all respondents as a whole.

Therefore, analysts should find the percent of respondents who answered free-response questions, and use that grain of salt when generalizing results across the population of survey respondents, or the population of the workplace, as a whole. Even if comments trend negatively, additional context provided by the quantitative questions in the survey may indicate that employees are generally content. So, responders’ bias should be accounted for.

With respect to topic modeling, there are fewer notorious limitations, but as with all clustering and classification techniques, there are likely to be comments or text that could fall into multiple topic bins. One sentence could address both leadership and training, so decisions need to be made about whether to dual-classify that comment; or associate it with whichever topic it more strongly associates with. Depending on the methodology, whether one is using Support Vector Machines, bottom-up or top-down hierarchical clustering methods, or otherwise, the groupings of the comments by topic may look different. Therefore, just like for sentiment analysis techniques, these methods are less accurate and less applicable for comments at the individual level, and should be restricted to generalizing insights at a higher, more summative level.

The first aspect on which this research focuses is bringing value to the sentiment classifications that the reviewers manually brought to the data. The goal of this portion was to identify the accuracy of existing sentiment analysis methodologies, and to attempt to enhance those that initially performed well, assessing accuracy by association to the manual classifications. In this section, the comments were run through various sentiment analysis methodologies, each utilizing different algorithms and drawing from different training sets. Scores were then scaled, and compared against the manual classifications to assess accuracy and variability. These results are reproducible and did not use any degree of randomness.

Table 1: Sentiment analysis techniques and respective details.

Technique	Lexicon options	Published (Naldi, 2019)	Package utilized for method in R	Polarity Scores
BoW with DTM	Syuzhet, AFINN, Bing, NRC	2015	Syuzhet	Weights may reflect intensity of sentiment
BoW with heavy pre-processing	Hu & Liu	2017	Meanr	Weights reflect classification of sentiment
Valence shifters, adversative conjunctions and DTM	Modified combination of Syuzhet and Hu & Liu	2016	sentimentr	Weights may reflect intensity of sentiment
StopWord Removal with heavy pre-processing and BoW Ratio	QDAP, multitude of other options	2017	SentimentAnalysis	Weights may reflect intensity of sentiment

In order to properly compare the performance of each sentiment analysis function, sometimes the same method was with different parameters, and the scores were scaled to mimic those of the manual review classifications. This meant that continuous scores were binned into “positive”, “neutral” and “negative”.

3.2 BoW with DTM

First, the Bag of Words with Document Term Matrix Method was used to calculate polarity scores for each of the comments in the data. This method is largely lexicon-based, a common approach explained above in Literature Review, and gives users the option to choose the lexicon they want to use. The Syuzhet Package in R was used to implement this methodology.

Essentially, this method takes a bag-of-words approach aided by a document-term frequency matrix. The bag-of-words approach separates the entire document (or, in this case, comment) into a list of words, and then computes a matrix identifying those words that appear next to one another. Without the document term matrix (DTM), there would be no remaining data indicating the structure of the document, for example, if a negator preceded a positive word such as “not happy”.

However, without the additional context that a lexicon provides, the function would not have an idea as to the weight, or perceived negativity or positivity, of a word such as “happy”. This is why there are so many different lexicons available, each developed and trained for different purposes. While the word “happy” is easy to interpret in any context, other words are very context-dependent. For example, the word “faded” may have different connotations depending on if it is describing denim jeans (i.e. positive) or antique furniture (i.e. negative). For each word, depending on the lexicon and the metric used, a polarity is associated, indicating the typical sentiment context in which that word is expressed.

The Syuzhet lexicon was developed by analysts in the Nebraska Literary Lab and ranges from -1 to 1 (Naldi, 2019). The AFINN lexicon began with a set of obscene words developed from Twitter and expanded to include over two-thousand words, including acronyms, and ranges in score from -5 to 5 on a continuous scale (Naldi, 2019). Finally, the Bing lexicon ranges from -1 to 1 and was developed by Minqing Hu and Bing Liu (Naldi, 2019).

Table 2: Information about applied lexicons.

Lexicon	Number of Words	Number of Positive Words	Number of Negative Words	Range	Type of Polarity Score
Syuzhet	10748	3587	7161	-1 to 1	continuous
AFINN	2477	878	1598	-5 to 5	discrete
Bing	6789	2006	4783	-1 to 1	binary
Hu & Liu	5787	2005	3782	-1 to 1	binary
Loughran-McDonalds (LM) Financial Dictionary	2709	354	2355	-1 to 1	binary
QDAP	4232	1280	2952	-1 to 1	binary
GI	3642	1637	2005	-1 to 1	binary
HE	190	105	85	-1 to 1	binary

The BoW with DTM Method was applied to the data with respect to three different lexicons, noted in Table 2. These lexicons determined the polarity scores of the words contained within them, and thus had different effects on the sentiment classifications when applied to the data. Summative results of these scores are detailed in the Results section of this paper, along with an assessment of accuracy when compared to the

manual sentiment classifications. Because the manual sentiment classifications were discrete, and some of these results are on a continuous or different discrete scale, all results were scaled to match that of the manual results. (Misuraca et al., 2020)

3.3 BoW with Pre-Processing

This methodology is much simpler than the previous BoW with DTM. Taking in a text string, this method is primarily focused on calculating polarity scores with term-level polarity aggregations. This is much less advanced than methods previously addressed, and leaves little to no room for customization. This research did not expect this method to perform particularly well in comparison with other, more advanced, methods. However, it was included to test notional assumptions about better methodologies, and this research noted that, when applied to larger datasets, analysts may benefit since the computing time may be significantly faster than other methods due to its simplicity and the ability to utilize parallel computing through the MeanR R Package (Naldi, 2019).

Essentially, taking in a text string, this method includes some pre-processing steps such as punctuation removal and removing capitalization. Then, for each word in the string, if the word appears in the lexicon (in this case, the Hu & Liu lexicon), then its associated polarity is assigned. If the word does not appear in the dictionary, it is assumed that the polarity is zero. Because the Hu & Liu lexicon is discrete, scores for each word are either -1, 0 or 1. Then, the score across the text is computed as the number of positively-scoring words minus the number of negatively-scoring words.

3.4 Valence Shifters and Adversative Conjunctions with DTM

This method further builds on and develops some of the concepts mentioned that may improve the performance of a sentiment analysis algorithm. In addition to taking into account negators and amplifiers, it creates a new classification of words encompassing those considered “valence shifters”. These are words that affect the degree to which a word is emphasized or de-emphasized by the writer. It also takes into account “adversative conjunctions”. Therefore, the phrase “very happy” will receive a more positive score, and “not happy” a more negative score, than the word “happy” on its own would have obtained.

Even when applied to the same lexicons as previous methods the polarity scores will not necessarily be the same. While the BoW with DTM method, for example, would recognize that the words are next to one another, rather than realizing one emphasizes or describes the other, it would treat both words individually according to their polarity score in the lexicon. In this method, rather than “not” and “happy” being treated individually, “not” is instead used to modify, or in this case reverse, the intensity of the polarity of “happy”.

In order to implement this method, the sentimentr package was used (Naldi, 2019). It reads in strings of text as character vectors, and uses punctuation characters to split the string into sentences. The analyst can specify the range that a valence shifter is able to affect. A range of 4 means that in the phrase “not happy, satisfied, or fulfilled”, “not” would be able to affect all three of the adjectives that follow it and shift those polarities, out to 4 words before or after. A range of 1 means that only the polarity of the word “happy” would be affected by the presence of “not”.

Due to the inclusion of malleable valence shifters dictionaries, this method is able to calculate the polarity of text strings not by summing term-level polarity scores or taking the ratio, but considering the words in the context in which they are present. For this reason, the study expected this method to perform better than those previously discussed.

The scoring methodology in the Valence Shifters and Adversative Conjunctions Method computes scores on a by-sentence basis. Therefore, to generate scores of a group of sentences, the individual scores are averaged and weighted by the word count in each sentence. This risks neutral sentences down-weighting a piece of

text, i.e. pulling positive sentence polarity scores down and negative polarity scores up. However, removing the neutral-scoring sentences would disrupt the continuity the study hopes to achieve by comparing metrics across methods, and so the averaging function was left as-is.

Therefore, to calculate comment-level polarity scores with this method, this research used simple weighted averaging (Fuchs, 2020; Raja, 2017).

3.5 Stopword Removal with Pre-Processing and BoW

The Stopword Removal with Pre-processing and BoW method was more recently developed, and was introduced as a concise SentimentAnalysis R Package in 2019 (Naldi, 2019). It also has the ability to draw from many more lexicons than previous methods addressed in this research. This is useful not only because of the sheer number of lexicons available, but also because of their contextual diversity. The lexicons available for this method include the Loughran-McDonald's Financial Dictionary developed in 2011 and the Qualitative Data Analysis Program (QDAP) dictionaries developed in 2019 (Naldi, 2019).

Unlike some methods previously discussed, this one does not generate scores as the algebraic sum of polarity scores per word or term. Instead, the default score is a ratio of the positive and negative terms. However, this can be changed depending on analyst preferences.

For this research, several methods were used to identify statistically significant words, in part to see the degree of variability between them and assess whether they indicated any useful insights as to topic-level respondent opinions. Additionally, if any words are context-dependent in the Air Force data and seem to be misinterpreted, and they appear in a list of statistically significant words, that would indicate that the accuracy of the function is way off and could be improved if the polarity of that word, or a set of related words, is adjusted.

3.6 Stopword Removal with Pre-Processing and BoW

This hypothesis will be tested by way of the research as explained in a study titled "Employee Pronoun Use In Verbatim Comments As A Predictor Of Job Attitudes And Turnover Intentions", published in 2014 through Wayne State University (Sund, 2017). For each comment, this research will count a total number of "we" and "non-we" pronouns, and calculate the percentage that this accounts for in the total words used.

In organizational psychology, this is called relationship literature, and pulls from the notion that pronouns of the "we" type indicate that the writer feels a sense of unity and community with their workforce and peers, while the use of "non-we" pronouns may indicate distancing and dissatisfaction between the author and their workforce (Slatcher & Vazire, 2008). For this research, a correlation matrix will then be created as demonstrated in the previously referenced research, and an ANOVA table will be used to assess the degree to which sentiment scores are correlated with pronoun usage.

Because the polarity of pronouns were said to influence the study, the first attempt at improving the performance of the SentimentAnalysis package was to adjust the stopword removal in the pre-processing phase to allow certain pronouns to remain in the document term matrix. Then, the SentimentLM lexicon, which did not currently contain any of these pronouns, was amended to include them, with either strictly negative or strictly positive weights associated with them since the SentimentLM lexicon is a binary dictionary. Several combinations of including pronouns in the positive and negative dictionaries were attempted, and the best combination for improving overall accuracy seemed to be weighting "you" and "they" pronouns negatively.

3.7 Context Word Adjustment

A similar approach was attempted to improve the performance of the dictionary with respect to Air Force specific words. While one may notionally be able to identify words that they may assume are context specific, if the word does not have a polarity at all, it will not be swinging the scores in the wrong direction, rather, it just will not contribute. However, words that are incorrectly classified may have a much bigger impact on the model performing poorly.

Therefore, to identify those at-risk words, the LASSO method was used in conjunction with regression analysis to identify words that contribute more heavily to the scores in the model. LASSO stands for least absolute shrinkage and selection operator), a regression analysis method originally formulated for application to linear regression models in attempts to improve prediction accuracies and model interpretability. Ordinary least squares (OLS) or generalized linear models (GLM) could have also been used.

Table 3 shows the list of words generated when compared to a variety of scores. This was not only run with the Manual, accurate scores, since for an analyst to identify at-risk words, they may not always have access to those manual scores. So, the study wanted to identify whether similar words appeared with other scoring mechanisms. Words have been stemmed by the pre-processing step, which is why some may look different than the direct terms. The four methods that had the highest strict accuracy scores thus far were investigated in this manner.

Table 3: Words identified by LASSO method as statistically significant, by scores.

Manual scores	**BoW with DTM (Bing Lexicon)**	**Valence Shifters and Adversative Conjunctions w DTM**	**Stopword Removal and Pre-processing with BoW Ratio (Sentiment HW Lexicon)**	**Stopword Removal and Pre-processing with BoW Ratio (Sentiment LM Lexicon)**	**Stopword Removal and Pre-processing with BoW Ratio (Sentiment HW Lexicon) with Pronoun Adjustment**
Intercept: -0.3487515 -0.04 peopl -0.03 posit -0.03 get -0.02 award -0.02 employe -0.02 work -0.01 need -0.01 opportun 0.01 job 0.02 train	Intercept: 0.5857396 0.01 train 0.03 get 0.03 level 0.06 time 0.07 leadership 0.07 peopl 0.10 employe 0.11 job 0.12 need 0.13 opportun 0.22 work 0.26 award	Intercept: 0.1696322 -0.01 get -0.01 leadership -0.01 peopl 0.01 level 0.01 award 0.04 opportun	Intercept: 0.04444311 0.01 posit 0.02 opportun	Intercept: 0.0430529 -0.01 need 0.01 posit 0.01 opportun 0.02 leadership	Intercept: -0.001219411 -0.20 get -0.16 job -0.12 posit -0.10 work -0.09 peopl -0.05 need -0.04 time -0.03 train -0.02 award -0.02 employe -0.01 level 0.05 opportun 0.22 leadership

As one looks across Table 3, note that outside of the first column, the methods are ordered by total accuracy percentage. One can see that the words identified as statistically significant in their contribution to the scores is similar across the different columns of the table. Statistically significant words with coefficients effectively at zero were not included in this table. These were tested at a 0.05 significance level. Coefficients

which are negative contributed to a negative weight when computing the sentiment scores, whereas coefficients which are positive contributed to a positive weight when computing the sentiment scores.

However, some words have negative coefficients in their contribution to the score, and have positive coefficients in their contribution to a different score. “Peopl” is always negative and shows up in each column. The word “get” is in each column, but is sometimes positive and sometimes negative. Words with a wider span and larger coefficient likely contribute more to the variability between the models. However, none of these seem to be largely context-dependent.

Regardless, the study found that in the SentimentLM dictionary, the words “posit” and “opportun” were in the dictionary as positive words. Therefore, an attempt was made to remove those words from the dictionary, since they are subjects of the question and thus should not have a polarity associated with them, and a new accuracy score was calculated. The results of that attempt are explained in Conclusions.

4.0 CONCLUSION

In the subsequent tables below is information reflecting the results of the applied methods with the comments in the data. Comparison between each method’s performances can be found in Table 6.

Queries about the code behind these methods can be found in the respective package libraries, which explain in detail the functions contained within packages and the arguments that the use may pass to those functions.

4.1 BoW with DTM Method Results

This method was applied with four different lexicons: Syuzhet, Bing, AFINN and NRC. As indicated in Table 4, the Syuzhet polarities ranged from -3.25 to 15.90, the Bing polarities ranged from -7 to 12, the AFINN polarities ranged from -13 to 36, and the NRC polarities ranged from -6 to 17. Table 4 indicates the range of the polarity scores at the comment level, and the frequency per bin when scaled for comparison with the manual scores.

Figure 1 indicates the distribution of the scores for each of the methods. These show that Bing was more centered on 0 while Syuzhet and NRC had longer tails into the positive scores. The distribution of the discrete scores from the manual reviewers is shown in Table 4 and Figure 1, with the number of negative scores being much higher than the number of positive scores and the number of neutral scores. In scaling the scores obtained here, scores greater than 0 are classified as positive, less than zero as negative, and zero as neutral. These are revisited for comparison in the Summaries section.

Table 4: BoW with DTM method lexicon results.

Lexicon	Min Polarity	Max Polarity	Average
Syuzhet	-3.25	15.90	1.69
Bing	-7.00	12.00	1.00
AFINN	-13.00	36.00	3.04
NRC	-6.00	17.00	2.05

**Polarity scores rounded to 2 decimal places*

4.2 Valence Shifters and Adversative Conjunctions Method Results

The Valence Shifters and Adversative Conjunctions Method was run over the vector of comments from the original data. There are options in the dynamic parameters to customize the valence shifter dictionary and the number of terms surrounding the valence shifter that it may affect.

The analyst may also alter the dictionary, as well as the weights of the valence shifters. As noted in methodology, the function was run once with the downweighted averaging function, and once with the average mean.

As one can see in Table 5 and Figure 1, the averaging method did not make a significant difference when compared to the down-weighted zeros, and so only scores from the first row method will be used for comparison.

Table 5: Valence shifters and adversative conjunctions method results.

Averaging Method	Min Polarity	Max Polarity	Average Polarity Score
average_mean	-1.125	2.221	0.1839
Default: down weighted zeros	-1.125	2.221	0.1816

4.3 Comparison Results

Figure 1 shows correlation matrices indicating the results of the initial sentiment analysis methods. The correlation indicates similarity or dissimilarity of word polarity between method and manual scores.

Manual scores are not strongly associated with any of the automated methods, with the highest absolute correlation (via the Pearson statistic) being 0.177 with the valence shifters approach. However, several of the automated methods are highly correlated. For example, the BoW with DTM Method, with Bing lexicon, and the Pre-processed BoW method, are very highly correlated, with r = 0.95. Figure 1 shows correlation of scores when binned for direct comparison to the manual scores.

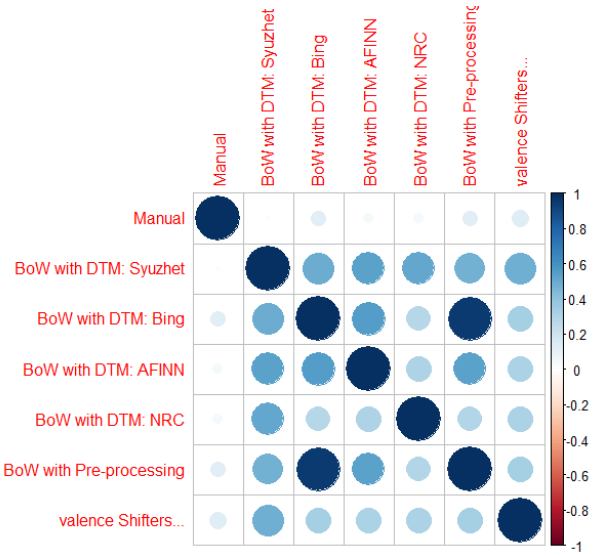

Figure 1: Correlation table of results from automatic methods.

Figure 1 and Table 6 indicate associations between the methods tested thus far beyond correlation, investigating degrees of accuracy. Equations for each of the columns are explained below.

Table 6: Method accuracy results.

Method	Total Accuracy	Positive Accuracy	Negative Accuracy	No Neutral Accuracy	OneSentiment Accuracy
BoW with DTM (Syuzhet)	0.227	0.069	0.948	0.208	0.226
BoW with DTM (Bing)	0.306	0.094	0.906	0.223	0.306
BoW with DTM (AFINN)	0.302	0.116	0.890	0.238	0.304
BoW with DTM (NRC)	0.232	0.057	0.869	0.185	0.229
Pre-processing with BoW	0.295	0.090	0.906	0.219	0.295
Valence Shifters and Adversative Conjunctions with DTM	0.307	0.225	0.942	0.339	0.308

***all scores rounded to 3 decimal places*

Total Accuracy is the sum of the correctly classified positive comments, the correctly classified negative comments, and the correctly classified neutral comments, divided by the total number of comments being classified.

Total Accuracy = (# correct Positives + # correct negatives + # correct neutrals) / total

Positive Accuracy is the sum of the correctly classified positive comments divided by the total number of manually-classified positive comments.

Positive Accuracy = (# correct Positives) / total True Positives

Negative Accuracy is the sum of the correctly classified negative comments divided by the total number of manually-classified negative comments.

Negative Accuracy = (# correct Negatives) / total True Negatives

No Neutral Accuracy is an attempt to remove from consideration those comments whose aggregate manual sentiment classification may be incorrect. Therefore, this calculates the performance of the algorithm when the

manually-classified neutral comments are removed, and it is the sum of the correctly classified positive and negative comments, divided by those comments which were manually assigned a sentiment other than neutral.

NoNeutral Accuracy = (#correct Positives + #correct Negatives) / (total - True Neutrals)

However, the Total Accuracy percentage may be perceived to be skewed due to the way in which the manual sentiment scores were classified for comments that expressed multiple sentiments. When those 97 comments manually classified as neutral are removed, the accuracy percentages shift, and the distribution of the classifications are instead 984 negative, 718 neutral and 171 positive. However, the manual scores became less correlated with the computer-generated scores, and the accuracy results did not significantly improve.

This research did not choose to remove the manually-classified neutral comments from consideration since it would not be realistic for a team of analysts to remove all those comments containing multiple sentiments from consideration without first having read through them.

One Sentiment Accuracy is calculated as the total accuracy for comments that only expressed one sentiment according to the manual review team. Therefore, comments that expressed multiple sentiments, and were thus averaged for the total manual score, were removed. This is to investigate the degree to which that manual sentiment method affects the accuracy scores. These are computed using the scaled results.

Table 6 indicates that the algorithms are very good at correctly classifying negative comments, but have low accuracy classifying positive comments. The effect is to drive down the accuracy of the package as a whole (Silge & Robinson, 2021)

4.4 Stopword Removal with Pre-Processing and BoW Method Results

Up to this point, the method performing best accounted for valence shifters and adversative conjunctions, which had an accuracy of 30.7% with the manual scores. Stopword removal with pre-processing and BoW were adjustments added in an attempt to improve accuracy. Table 7 shows the results from one run of the method, drawing from 4 different lexicons (Feuerriegel & Proellochs, 2019; Misuraca et al., 2020).

Table 7 shows this enhanced method performs better in total accuracy and better in correctly classifying positive comments when compared to the previous methods. However, the accuracy for correctly classifying negative comments is significantly lower.

Table 7: Lexicons for valence shifters and adversative conjunctions method.

Lexicon	Total Accuracy	Positive Accuracy	Negative Accuracy	No Neutral Accuracy
SentimentGI	0.217	0.080	0.911	0.211
SentimentHE	0.308	0.036	0.534	0.115
SentimentLM	0.374	0.207	0.675	0.281
SentimentQDAP	0.236	0.050	0.932	0.190

***all scores rounded to 3 decimal places*

Table 8 shows the results of the Pronoun improvement attempts on this method with the SentimentLM lexicon as detailed in Methodology.

Table 8: Accuracy results with pronoun adjustment.

Method	Total Accuracy	Positive Accuracy	Negative Accuracy	No Neutral Accuracy
SentimentLM	0.374	0.207	0.675	0.281
You added to Negative	0.451	0.694	0.182	0.435
You and They added to Negative	0.471	0.678	0.180	0.479
You and I added to Negative	0.454	0.688	0.180	0.442
You: Negative We: Positive	0.451	0.693	0.182	0.435
You, They and I Negative	0.446	0.694	0.174	0.414

An additional attempt to improve the performance of this method was to remove those words that were subjects of the question posed to respondents and listed as significant contributors to the score as a whole, and in the lexicon from which the function was pulled. The result was the removal of two stemmed wordsm "posit" and "opportun". After removal from the dictionary, the scores were recalculated and the accuracy re-assessed. displays the results. Even in conjunction with the pronoun rule previously implemented, and currently with the highest total accuracy, this did not improve that approach to any significant degree.

Ultimately, the best-performing model was the Valence Shifters and Adversative Conjunctions Method with the SentimentLM lexicon, amended with the pronouns dictionary. Table 9 summarizes the results of the comparison between method scores and the manual assessment.

Table 9: Accuracy results with context word adjustment.

	Total Accuracy	Positive Accuracy	Negative Accuracy	No Neutral Accuracy
Words pulled from SentimentLM	0.414	0.712	0.176	0.363
Words pulled from SentimentLM & Pronoun Approach	0.471	0.678	0.180	0.479

4.5 Implications

All results drawn from this analysis are only directly applicable to the specific context of the survey data on which the methodologies and models were tested. That survey data is centric to the Financial Management civilian career field in the United States Air Force.

This research indicates that automated sentiment classification techniques are insufficient in garnering the sentiment of a piece of text as a whole when applied as the sole classification method. Instead, analysts should pair sentiment classification techniques with careful parameters which better suit the algorithm to the context to which it is being applied. This may mean implementing one or more of the adjustment techniques explored in this research, building a lexicon for the analysis use-case specifically, or training an algorithm on a smaller set of manually classified, taken-as-truth, classifications to the set. In this research, the truth data metric, i.e. the manually assigned classifications, may not have been nuanced enough to allow for sufficient comparison to the automated techniques. Manual classifications calculated in a different manner may yield different results.

Additionally, the topic modeling algorithms seemed promising in their application to the data, when the attempt was made to use that algorithm to sort answers to all of the questions into the main themes addressed. Therefore, it may be a viable recommendation that analysts first identify the topics which they wish to further investigate in a dataset of textual comments, and then apply sentiment analysis as a secondary technique, thereby parsing the data into a smaller set and isolating those comments which address a given topic. If techniques are applied in this order, not only can the analyst then apply sentiment analysis techniques specifically to a unique lexicon, whose use-case is clear, but it may also be more viable to visually check the performance of a sentiment analysis classification against those comments to which it is being applied.

Any assumptions about degrees of confidence in the tangential application of these techniques to other United States Air Force career fields, other organizations in the government or Department of Defense, or parallel career fields for active duty and enlisted employees, should be carefully verified. While this research is intended to bring a greater degree of understanding to linguistic analysis in an Air Force context, implications about expected degrees of accuracy should be verified when applied to new contexts.

4.6 Recommendations for Further Research

Future research could include further exploration of some of the techniques identified as more effective, with higher accuracy. Analysts may also explore the degree to which a lexicon developed specifically for application to a Department of Defense or Air Force specific context may perform when compared with generalized lexicons. More study is needed to identify the aspects of this field of study which influence accuracy percentages and the performance of sentiment analysis models. Aspect-level sentiment analysis, even when paired with flawless topic modeling, would not perform well given the lack of confidence in sentiment analysis results, even at the sentence level, for either specific identification or generation of summative numbers.

5.0 REFERENCES

[1] A., V., & Sonawane, S. S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, 139(11), 5-15. https://doi.org/10.5120/ijca2016908625

[2] Almatarneh, S., & Gamallo, P. (2019). Comparing Supervised Machine Learning Strategies and Linguistic Features to Search for Very Negative Opinions. Information, 10(1), 16. https://doi.org/10.3390/info10010016

[3] Araujo, Diniz, & Bastos. (2016, March 31). iFeel 2.0: A Multilingual Benchmarking System for Sentence-Level Sentiment Analysis. 10th International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13039

[4] Chen, E. (2011). Introduction to Latent Dirichlet Allocation. Retrieved January 18, 2021, from https://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

[5] Clark, M. (2018). An Introduction to Text Processing and Analysis with R. Retrieved December 22, 2020, University of Michigan. https://m-clark.github.io/text-analysis-with-R /

[6] Council. (2020, December 16). Council Post: 11 Ways (And Reasons) To Measure The ROI Of Your Company Culture. Forbes. https://www.forbes.com/sites/forbeshumanresourcescouncil/2020/12/16/11-ways-and-reasons-to-measure-the-roi-of-your-company-culture/

[7] Feuerriegel, S., & Proellochs, N. (2019, March 26). Sentiment Analysis Vignette [Cran.r-project.org].

[8] Fuchs, M. (2020, December 28). Doing your first sentiment analysis in R with Sentimentr. Medium. https://towardsdatascience.com/doing-your-first-sentiment-analysis-in-r-with-sentimentr-167855445132

[9] Huselid, M. A. (1995). The Impact of Human Resource Management Practices on Turnover, Productivity, and Corporate Financial Performance. Academy of Management Journal, 38(3), 635-872.

[10] Jagtap, V.S., and K. Pawar. "Analysis of Different Approaches to Sentence-Level Sentiment Classification." International Journal of Scientific Engineering and Technology, vol. 2, no. 3, 2013, pp. 164-170.

[11] Jipa, G. (2019). The Value of Structured and Unstructured Content Analytics of Employees' Opinion Mining. Journal of Administrative Sciences and Technology, 2019, 1-25. https://doi.org/10.5171/2019.908286

[12] Judd, S. (2018, March 14). Employee Surveys Are Still One of the Best Ways to Measure Engagement. Harvard Business Review. https://hbr.org/2018/03/employee-surveys-are-still-one-of-the-best-ways-to-measure-engagement

[13] Kenny, G. (2020, September 14). What Are Your KPIs Really Measuring? Harvard Business Review. https://hbr.org/2020/09/what-are-your-kpis-really-measuring

[14] Koch, K. (2020, March 26). A Friendly Introduction to Text Clustering. Medium. https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04

[15] Luo, N., Zhou, Y., & Shon, J. (2016). Employee Satisfaction and Corporate Performance: Mining Employee Reviews on Glassdoor.com. 16. 37th International Conference on Information Systems, Dublin, Ireland

[16] Misuraca, M., Forciniti, A., Scepi, G., & Spano, M. (2020). Sentiment Analysis for Education with R: packages, methods and practical applications. 27. arXir: 2005.12840

[17] Naldi, M. (2019). A review of sentiment computation methods with R packages. ArXiv:1901.08319 [Cs]. http://arxiv.org/abs/1901.08319

[18] Pröllochs, N., Feuerriegel, S., & Neumann, D. (2018). Statistical inferences for polarity identification in natural language. PLOS ONE, 13(12), e0209323. https://doi.org/10.1371/journal.pone.0209323

[19] Raja, A. (2017). Handling 'Happy' vs 'Not Happy': Better sentiment analysis with sentimentr in R. DataScience+. Retrieved January 29, 2021, from https://datascienceplus.com/handling-happy-vs-not-happy-better-sentiment-analysis-with-sentimentr-in-r/

[20] Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). SentiBench-A benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1), 1-29. https://doi.org/10.1140/epjds/s13688-016-0085-1

[21] Salas-Zárate, M. del P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017, February 19). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach [Research Article]. Computational and Mathematical Methods in Medicine; Hindawi. https://doi.org/10.1155/2017/5140631

[22] Sund, A. E. (2017). Employee Pronoun Use In Verbatim Comments As A Predictor Of Job Attitudes And Turnover Intentions. 71.

[23] Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications, 34(4), 2622-2629. https://doi.org/10.1016/j.eswa.2007.05.028

[24] Wang, N. (2017, November 21). Topic modeling and sentiment analysis to pinpoint the perfect doctor. Medium. https://blog.insightdatascience.com/topic-modeling-and-sentiment-analysis-to-pinpoint-the-perfect-doctor-6a8fdd4a3904