Comparison of LIWC and Machine Learning Model in Predicting Rumination

Logan Delgado, B.S.

Georgia Southern University

Abstract

Rumination may conceptually be defined as a negative thinking process. It is considered a strong risk factor in the development of depression. Furthermore, research has suggested that the role of rumination is linked to multiple psychopathologies including OCD, eating disorders, post-traumatic stress disorder, and alcohol dependence. It is a strong disruptor in normal, daily functioning. Due to the transdiagnostic role that rumination plays in psychopathology, it is important to identify early manifestations before advanced psychopathology becomes apparent. One way to identify early manifestations is through text. Text analysis tools, such as LIWC and machine learning model, Support Vector Machine (SVM), are to be used in analyses. The goal of the present research project is to compare the psychological software, LIWC, with the sentiment analysis model, 'SVM', to see if either text analysis better predicts the manifestation of rumination in samples of journal prompts.

Introduction

Rumination has been a difficult concept to define due to researchers' inability to identify the underlying mechanisms of rumination (Sansone & Sansone, 2012). Therefore, a few definitions are illustrated below. Watkins & Robert (2020) posit that rumination may be understood as a repetitive, prolonged, and recurrent negative thinking process concerning feelings, experiences, and worries pertaining to the self. Rumination, according to Sansone & Sansone (2012), is a "detrimental psychological process characterized by perseverative thinking around negative content that generates emotional discomfort" (p.2). Rumination acts as a disruptor when an event or experience occurs in which the negative thought processes start to unravel for the individual. Research suggests the role of rumination could be found in psychotic experiences, vulnerability and outcome in eating disorders, and exacerbation of alcohol dependence (Luca, 2019). Research has shown that this maladaptive coping and reflective mechanism plays a significant role in the development, duration, and maintenance of psychopathologies and impairment in normal functioning (Luca, 2019). Going back to the conceptualization of rumination at the beginning of this paragraph emerged a theme. Rumination is a tendency to focus on the negative content or overarching negative thinking process. With this

in mind, it may be helpful to consider rumination and its association with negative affect or thinking. Negative mood and experiences helps in the persistence of cognitive rumination (Nolen-Hoeksema et al., 2008). Furthermore, if an individual with a negative self-concept (e.g. "I am bad") was asked to write about anything for 20 minutes (such as this journal prompts participants to), then the individual may be triggered into ruminative thinking. This relates to text analysis because this ruminative thinking and negative self-concept could lead to use of negative words in their journal prompt. Due to the transdiagnostic role of rumination in multiple psychopathologies and its association with negativity in affective and thinking processes, it is likely that cognitive rumination may manifest through negative emotionality in text. Furthermore, the use of negative emotionality in text may be something tangible for text analyses models to capture.

LIWC

A popular and notable facet of psycholinguistics is the LIWC software (Pennebaker et al., 2015). The LIWC operates as a word count frequency and classification word-program. LIWC has been utilized in text analysis extensively in psychological research, but its accuracy in predicting rumination manifestation is unknown. LIWC operates as such: assigns a word to a particular category, as it either meets the category or it does not (binary fashion). LIWC contains a *Negative Emotion* domain most pertinent to this research project (Pennebaker et al., 2015). For the *Negative Emotion* category, words are placed into this category and this category will be the basis for the accuracy of predicting rumination. For example, LIWC operates by working as a dictionary or list of words in the negative emotion feature. When LIWC is analyzing text and notices a word that is in its negative text dictionary, that word is flagged and subsequently given a score of "1" as an indication it meets criteria as a negative emotion word. This research is

comparing the classification: “Negative Emotion” with scores in rumination against the sentiment analysis model. The LIWC software program may not perform as well as its counterpart, machine learning model, in predicting rumination (Cutler et al., 2020).

Support Vector Machine

Sentiment analysis has gained more widespread attention and research in recent years pertinent to the field of psychological research. Sentiment analysis is the use of natural language processing, text analysis, computational linguistics and extracts emotional contexts transforming it into a quantitative analysis. Machine Learning is a subset field of sentiment analysis. Machine learning sentiment analysis refers to training a machine learning model to understand the polarity of the word order using a sentiment-labeled training set. Support Vector machines are a type of machine learning model classified in the supervised learning domain. Supervised learning refers to the process of priori training a model with labeled data in order to gain a more precise classification (Chaovalit & Zhou, 2005). To demonstrate, code will be run in order for the Support Vector Machine to operate in *R.* Once code is run, I will train the model by going through a certain percentage of data and classify the text as either rumination or not ruminative. My classification will be done by comparing the text with the self-report of the Rumination facet of the Rumination-Response Questionnaire. It is important to discuss a few advantages and disadvantages of Support Vector Machine. A few advantages of this machine learning model is that it operates high dimensionality which is pertinent to this research project since it can be used for document classification and sentiment analysis, its memory efficiency is impressive, and its versatility as a model may prove to be more advantageous than LIWC. A few disadvantages of *SVM* include: Kernel Parameters Selection which may lead to poorer classification performance,

there is no probabilistic explanation for any classification done by this model, and it does not operate well for

Though LIWC and *SVM* are similar in their textual analysis process, LIWC is able to classify in a binary fashion whether a word meets dozens of the software's classifications. The author chose the *Negative Emotion* value in predicting rumination due to the rumination's emphasis on negative emotional affect and cognitive state. *SVM* does similarly operate in a binary classification fashion. The Support Vector Machine classifies text, in this project, as either a manifestation of rumination or not a manifestation of rumination. While LIWC assigns a word to a classification (e.g. *Negative Emotion*) and then this classification will be used to predict cognitive rumination; *SVM* does not operate like this. *SVM,* once finished training, will classify the text as either ruminative in nature or not ruminative. This accuracy will be compared with the *Negative Emotion* classification of LIWC.

Sentiment analysis may capture the role of rumination (as measured by its predictive success rate) as it is successful in predicting the manifestation of depression in social media posts and other settings (Babu & Kanaga, 2021; Chen et al., 2018). While there has been extensive work done to predict depression using semantic analyses models (Babu & Kanaga, 2021), research on semantic analyses predictive modeling on rumination is sparse. This machine learning sentiment analysis tool, *SVM,* will be more successful (i.e. success rate of at least 65%; Onan, 2018) in predicting the manifestation of rumination, than LIWC, as measured by self-reports of cognitive rumination and journal prompts (qualitative data).

Method

Data collection for this research project has already been completed, therefore, a power analysis will not be conducted to see how many participants are needed for a specific effect size.

This project will utilize archival data from three separate samples of research participants. This data is stored under the discretion of Dr. Nicholas Holtzman and collaborators. The total, valid number of participants from three separate samples conducted by Holtzman et al. is 1,939 participants. Participants were mostly female (n = 1,409 women) and white. Participants were recruited from SONA at Georgia Southern University and another University in the Northern Great Lakes region (Holtzman et al., unpublished manuscript).

For this research project, a sentiment analysis model will be calculated on qualitative text. This model can be found in the *R* programming language under the *tidytext* package.

Below is sample code to be used to access the *SVM* dictionary for analyses in the *R* programming language; this code was taken from DataCamp(2018). This code will need to be extensively modified in order to run sentiment analysis and classification for this research project.

```
install.packages("e1071")
library(e1071)
set.seed(10111)
x = matrix(rnorm(40), 20, 2)
y = rep(c(-1, 1), c(10, 10))
x[y == 1,] = x[y == 1,] + 1
plot(x, col = y + 3, pch = 19)
dat = data.frame(x, y = as.factor(y))
svmfit = svm(y ~ ., data = dat, kernel = "linear", cost = 10, scale = FALSE)
print(svmfit)
plot(svmfit, dat)
```

Participants completed demographic information (ex. Age, Gender, Race, etc.) and the 24-item Rumination-Reflection Questionnaire [RRQ; Trapnell and Campbell (1999). A sample item from the Reflection facet is “I love exploring my inner self.” and a sample item from the Rumination facet is “My attention is often focused on aspects of myself I wish I’d stop thinking about.” These items were assessed on a Likert scale from 1 “strongly disagree” to 5 “strongly

agree". This measure has been found to have good reliability ($\alpha = 0.92$ for rumination, $\alpha = 0.90$ for reflection) and convergent validity (Trapnell and Campbell, 1999).

For LIWC, the negative emotionality column will be emphasized and used for predicting rumination. For example, the word, "sad" would be given a score of 1 in the negative emotionality column indicating that sad represents negative emotionality. If a word was given a score of 0, then the word would not represent negative emotionality (Pennebaker, 2015).

For *SVM*, the model will be trained to decipher whether a text is ruminative. Once the model is trained to understand whether a text meets the rumination classification, it will be moved on to be evaluated to ensure no more improvements need to be made to the classification system of this machine learning model.

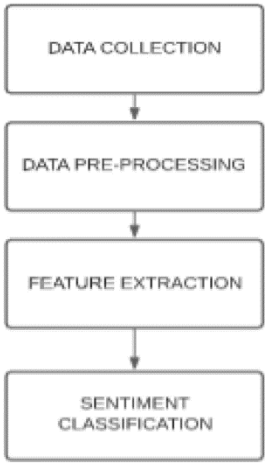
Participants signed up via SONA for course participation or extra credit. Data was collected via Qualtrics and SurveyMonkey (Holtzman et al., unpublished manuscript). Participants completed the materials previously discussed and moved on to the journal prompt where they were instructed to write for 20 minutes.

Participants were to respond to the following prompt: "For the next 20 minutes, write about whatever comes to your mind. Think about what your thoughts, feelings, and sensations are at this moment. Write about them as they come to you; follow where your mind naturally goes." (Holleran & Mehl, 2008). Once the 20 minutes had been completed in the journal prompt section, participants were allowed to finish the study.

LIWC data analysis was collected as part of the archival research (Holtzman et al., unpublished manuscript) so a regression analysis will be done in R programming language with the *Negative Emotion* column of LIWC to be used as a predictor with the dependent variable

being the *Rumination* facet of the Rumination-Reflection Questionnaire (Trapnell & Campbell, 1999).

Below is a figure showcasing the stages from data collection to sentiment classification for the SVM model. Since data collection has already ended, this research project is at the Data



Pre-Processing Stage (Stage 2; Babu & Kanaga 2021).

To ensure that this model does not overfit the data – essentially suggesting that the model is not biased towards a particular data set and is relevant to others – it is important for some of the data to be part of a training set, evaluation set, and the rest of the data to serve as the testing data. To clarify, the model will be evaluated against three data sets of the journal entries. The first step of the evaluation is to use a training set (AKA what data the model will be trained on – this will amount to 50% of the total valid journal entries) to tweak the mechanics (e.g. being able to detect sarcasm potentially and becoming more precise in detecting rumination) of the sentiment analysis tool. The next step is to evaluate prediction success against an evaluation set

(AKA a "test run" before the actual test – roughly 20% of the total valid journal entries) to ensure that no more improvements in the model are needed. The last step is to finally utilize the model on the testing set (AKA compare the final results of the model to the official data – roughly 30% of the total valid journal entries). This helps ensure that the model is being instrumentally improved and can be generalized to other data sets even if the data falls under the same scope.

Design

A multiple regression approach will be utilized to assess accuracy of LIWC and accuracy of the *SVM* sentiment analysis model in predicting rumination scores. The *Negative Emotion* values of LIWC and the negative sentiment of *SVM* will be used as predictors in the analyses to predict the Rumination facet of the Rumination-Response Questionnaire (outcome/dependent variable).

Fake Results

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.291[a]	.085	.065	1.35032	.085	4.351	1	47	.042
2	.690[b]	.476	.454	1.03247	.392	34.394	1	46	.000

a. Predictors: (Constant), Negative emotional facet of LIWC

b. Predictors: (Constant), Negative emotional facet of LIWC, Sentiment Analysis Model

c. Dependent Variable: Scores on Rumination Measure

To illustrate the success of the machine learning model compared to the LIWC, a linear regression model was used. These results show that the LIWC accounts for 8.5% of the variance

in Rumination scores. LIWC and the SVM model combined to account for 47.6% of the variance in rumination scores. This manifests in 39.2% of the variance in rumination scores being explained by the Machine Learning Sentiment Analysis (*SVM*) model that was not explained by the LIWC model. This gives further credibility that machine learning sentiment analysis should be widespread in text analysis in research pertaining to cognitive rumination.