Text Analytics Notes - Sentiment Analysis for Specific Emotions

Promothesh Chatterjee*

Sentiment Scores

A lot of packages give sentiment scores, let's take a quick look on how these are computed before exploring sentiment analysis using specific emotions.

A computer can't naturally understand words as they aren't directly calculable. Take a simple product review, for instance:

"The comforter was very comfortable. Good quality but $199 was a terrible price."

You can't simply add these words together. By removing numbers, punctuation, and common stop words, we get: "comforter very comfortable excellent quality terrible price." To evaluate this, we use a lexicon that assigns numbers to words: positive words = 1, negative words = -1, and neutral words = 0. This results in a numerical vector: 0 + 0 + 1 + 1 + 1 - 1 + 0 = 2. Since 2 is greater than zero, the review is considered positive. Another method to determine sentiment involves calculating a ratio, subtracting the number of negative words from positive words and dividing by the total number of words, represented by a formula.

Sentiment Score = (number of positive words - number of negative words) / total number of words Sentiment Score = (3 - 1) / 7 = 0.2857143

In this example, the sentiment score is 0.3. Here, the classification is binary—either positive or negative. Using this approach, both "bad" and "catastrophic" receive a score of -1, even though "catastrophic" is more negative than "bad."

Not all dictionaries, calculate sentiment in the manner of the formula (positive - negative) / total. Another, more nuanced method, assigns words a score upon a continuous scale, with words of greater extremity bearing higher values. Thus, "catastrophic" might be accorded the severe score of -0.90, whilst "bad" might bear the milder score of -0.30. With this approach, the sentiment scores of all words within a text are aggregated to divine the overall sentiment score.

Sentiment scores may be given upon words through diverse means. Experts may do so manually; alternatively, crowdsourcing platforms such as Amazon Mechanical Turk may gather the collective judgment of many. Moreover, the machine learning algorithms may automate this process. The lexicons thus manually wrought by experts are typically more precise, albeit smaller, due to the considerable time and cost required for their creation. Meanwhile, those lexicons generated by automatic means are more expansive, their precision contingent upon the efficacy of the algorithms employed.

Sentiment Analyses incorporating emotions

Most of the analyses we have seen so far computes the polarity of the text and reports it as positive or negative. Sometime we want more information than that. For example, we might want to know about the specific emotions contained in the text. One popular dictionary is called LIWC.

The Linguistic Inquiry and Word Count (LIWC) is a widely used text analysis tool developed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis. LIWC analyzes the emotional, cognitive, and structural components present in individuals' written or spoken language. Here's an overview of LIWC:

Purpose and Functionality

LIWC is designed to quantify various linguistic and psychological categories within a text. It is used in a range of fields, including psychology, sociology, linguistics, marketing, and health research, to understand and measure the underlying psychological states, emotional expressions, and cognitive processes of individuals based on their language use.

Key Features

1. **Linguistic Analysis**: LIWC analyzes texts based on a dictionary that includes thousands of words and word stems. Each word is categorized into multiple linguistic and psychological categories.
2. **Psychological Dimensions**: LIWC measures dimensions such as:
    - Emotional tone (positive or negative emotions)
    - Cognitive processes (e.g., insight, causation)
    - Social processes (e.g., family, friends)
    - Biological processes (e.g., body, health)
    - Personal concerns (e.g., work, leisure)
3. **Output Metrics**: LIWC provides percentages of words in each category, offering a comprehensive profile of the linguistic and psychological features of the text.

Categories and Dictionaries

LIWC dictionaries are comprehensive and include a wide range of categories. Some primary categories include: - **Function Words**: Pronouns, articles, prepositions, etc. - **Affective Processes**: Words related to emotions, such as happiness, sadness, anger. - **Cognitive Processes**: Words indicating thinking styles, such as cause, insight, certainty. - **Perceptual Processes**: Words related to sensory experiences, like seeing, hearing, feeling. - **Social Processes**: Words related to social relationships, like family, friends, humans. - **Biological Processes**: Words related to bodily functions, health, etc. - **Drives**: Words indicating motives, such as achievement, power, reward. - **Time Orientation**: Words indicating temporal focus, like past, present, future.

Applications

1. **Psychological Research**: Studying the relationship between language use and psychological states, such as mental health, stress, and coping mechanisms.
2. **Sociolinguistics**: Understanding how language reflects social identity, group dynamics, and cultural norms.
3. **Marketing and Consumer Behavior**: Analyzing customer reviews, social media posts, and other user-generated content to gauge consumer sentiments and preferences.
4. **Health Communication**: Examining the language used in patient narratives, medical records, and health communications to identify psychological and emotional states.
5. **Political Science**: Analyzing political speeches, debates, and social media to understand politicians' communication styles and the public's response.

Development and Validation

LIWC dictionaries have been developed through extensive research and validation. The dictionaries are updated periodically to include new words and reflect changes in language use.

Software and Usage

LIWC is available as software that can be used to analyze text data. Users input their text into the software, which then processes the text and provides output metrics based on the LIWC dictionary. The software is user-friendly and widely accessible for researchers and practitioners.

Example

In a study analyzing social media posts, researchers might use LIWC to determine the frequency of words related to emotions like "happy" or "sad" to understand the overall emotional tone of the posts.

LIWC is a powerful tool for anyone interested in exploring the intersection of language and psychology, providing deep insights into how language reflects and influences human thought and behavior.

The problem with LIWC is that you need a paid subscription to use it but is incredibly popular among researchers.

THE NRC LEXICON

The NRC Emotion Lexicon, often referred to as the NRC Dictionary, is a resource developed by the National Research Council Canada (NRC) for sentiment analysis and emotion detection in text. It contains lists of English words and their associations with eight basic emotions and two sentiments. Here's a detailed overview:

Emotions Covered

The NRC Emotion Lexicon identifies eight primary emotions: 1. **Anger** 2. **Anticipation** 3. **Disgust** 4. **Fear** 5. **Joy** 6. **Sadness** 7. **Surprise** 8. **Trust**

Sentiments Covered

The lexicon also includes associations with two primary sentiments: 1. **Positive** 2. **Negative**

Structure and Content

- **Word-Emotion Associations**: Each word in the lexicon is annotated with binary associations (yes/no) indicating whether it is associated with a particular emotion or sentiment.
- **Comprehensive Coverage**: The lexicon contains over 14,000 words, making it one of the more comprehensive resources for emotion and sentiment analysis.

- **Binary and Scored Associations**: Words are typically annotated with binary associations, but there are also extended versions with intensity scores indicating the strength of the association with a particular emotion or sentiment.

Applications

- **Sentiment Analysis**: Used to determine the sentiment (positive/negative) of a text, useful for customer feedback analysis, social media monitoring, etc.
- **Emotion Detection**: Helps in detecting the specific emotions expressed in text, valuable in fields like psychology, marketing, and human-computer interaction.
- **Natural Language Processing (NLP)**: Enhances the ability of machines to understand human emotions and sentiments in text, aiding in tasks like chatbots, recommendation systems, and content analysis.

Access and Usage

The NRC Emotion Lexicon is publicly available for research purposes. It is widely used in academic research, industry applications, and open-source projects. It can be accessed and downloaded from various repositories or directly from the NRC Canada website.

Development and Validation

The lexicon was developed through a combination of manual annotation and crowdsourcing. The development process included validation by multiple annotators to ensure accuracy and reliability.

Example Usage

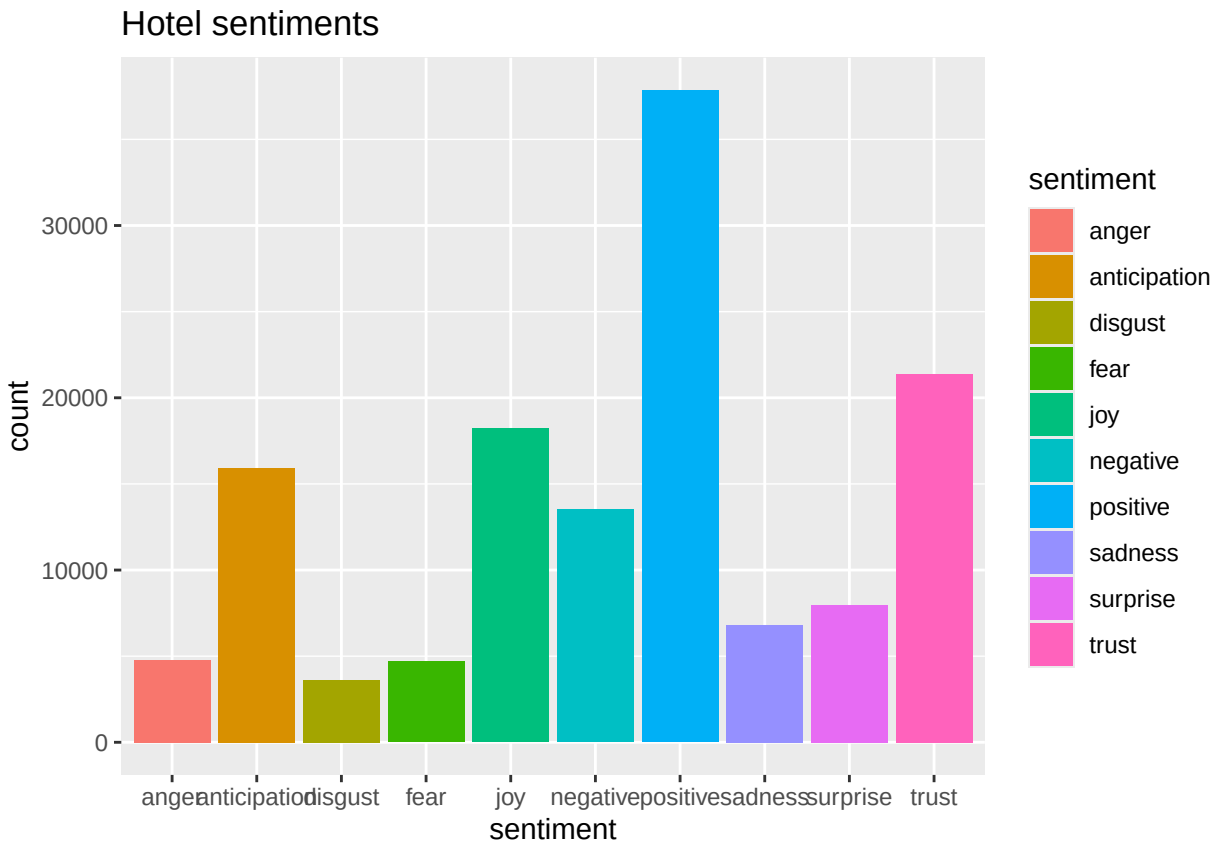
A typical use case in sentiment analysis might involve processing a batch of text data, tokenizing the text into individual words, and then using the NRC Emotion Lexicon to count the occurrences of words associated with each emotion and sentiment.

The NRC Emotion Lexicon is a valuable tool for anyone looking to incorporate emotion and sentiment analysis into their text processing workflows.

Using R for NRC lexicon

```
library(syuzhet)
library(tidyverse)

# Read in the data
hotel_raw <- read_csv("C:/Users/u0474728/Dropbox/Utah Department Stuff/Teaching/Text Analysis/Summer 20
```

```r
set.seed(122335)
# Take a small random sample of the data
hotel_raw <- slice_sample(hotel_raw,n= 5000)

# Compute NRC sentiment
nrc_data <- get_nrc_sentiment(hotel_raw$Description)
df_combined <- bind_cols(hotel_raw, nrc_data)

## Visualize Emotions

#transpose
td<-data.frame(t(nrc_data))

#The function rowSums computes column sums across rows for each level of a grouping variable.
td_new <- data.frame(rowSums(td[1:ncol(td)]))

#Transformation and cleaning
names(td_new)[1] <- "count"
td_new <- cbind("sentiment" = rownames(td_new), td_new)
rownames(td_new) <- NULL
td_new2<-td_new[1:8,]

#Plot One - count of words associated with each sentiment
quickplot(sentiment, data=td_new, weight=count, geom="bar", fill=sentiment, ylab="count")+ggtitle("Hotel
```


Hotel sentiments
count
30000
20000
10000
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust