Plan:

1. Define Text Analysis
2. Introduce Sentiment Analysis
3. Work through example of sentiment analysis

Text Analysis: Sentiment Analysis

Shannon E. Ellis, Ph.D
UC San Diego

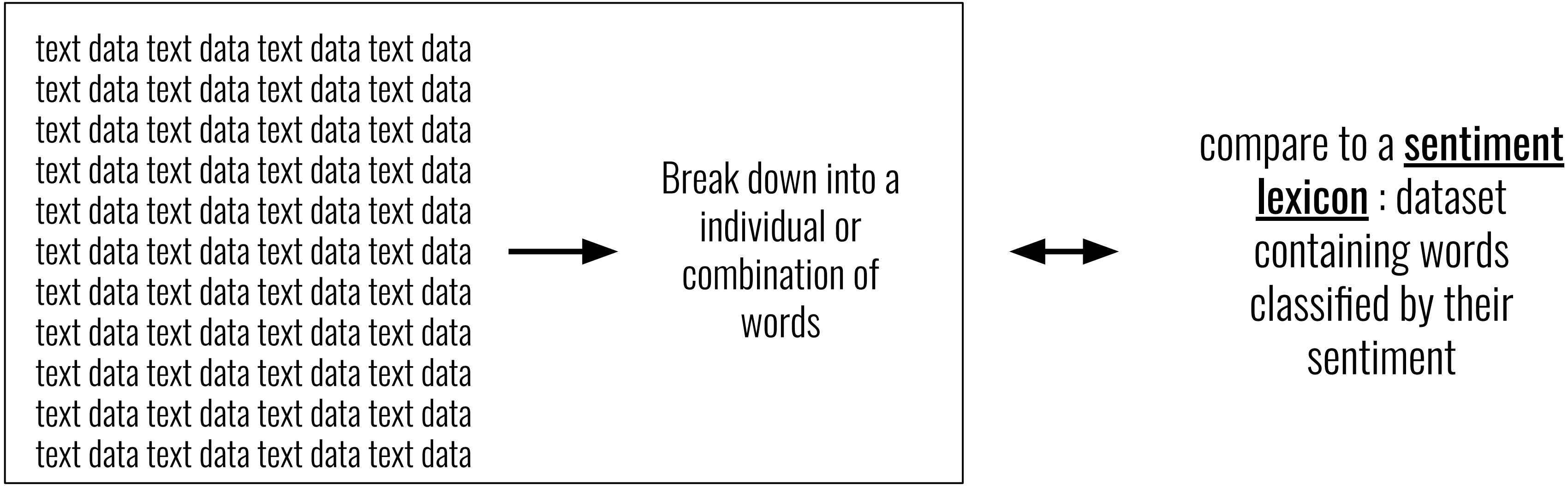
Department of Cognitive Science
sellis@ucsd.edu

Examples of questions that require text analysis

1. Did J.K. Rowling write The Cuckoo's Calling under the pen name Robert Galbraith?
2. What themes are common in 19th century literature?
3. Do the angriest tweets come from Trump himself?
4. Is Hillary the most poisoned name in US History?

Sentiment Analysis

Programmatically infer emotional content of text

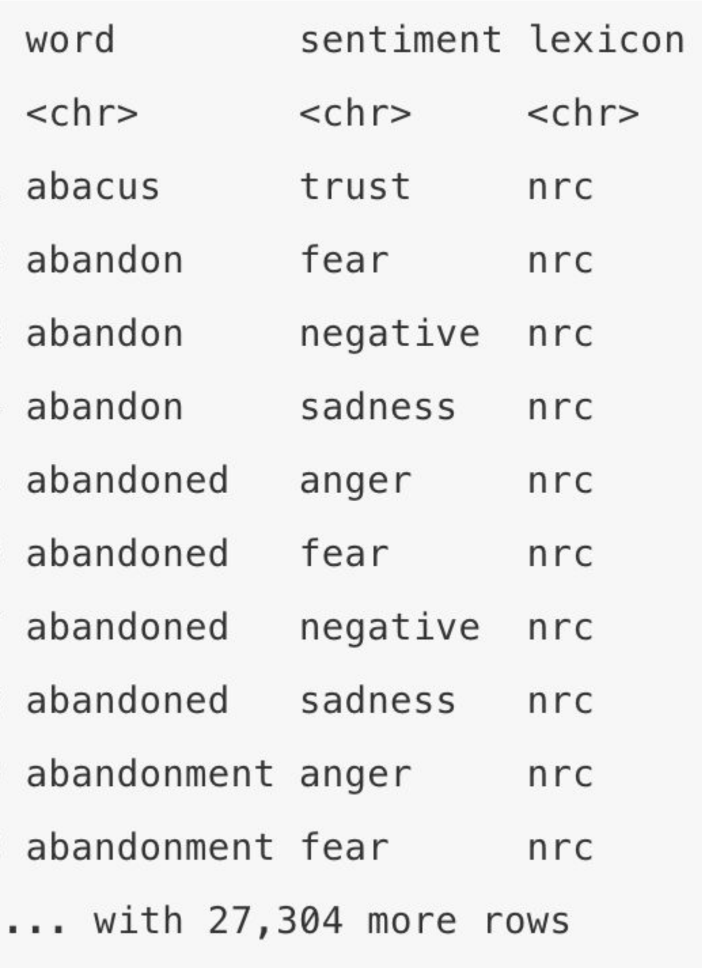


text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data

Break down into a individual or combination of words

compare to a **<u>sentiment lexicon</u>** : dataset containing words classified by their sentiment

Part of the “NRC” sentiment **lexicon**:



```
 word        sentiment lexicon
 <chr>       <chr>     <chr>
 abacus      trust     nrc
 abandon     fear      nrc
 abandon     negative  nrc
 abandon     sadness   nrc
 abandoned   anger     nrc
 abandoned   fear      nrc
 abandoned   negative  nrc
 abandoned   sadness   nrc
 abandonment anger     nrc
 abandonment fear      nrc
... with 27,304 more rows
```

http://varianceexplained.org/r/trump-tweets/

When doing sentiment analysis…

token - a meaningful unit of text

- what you use for analysis
- *tokenization* takes corpus of text and splits it into tokens (words, bigrams, etc.)

stop words - words not helpful for analysis

- extremely common words such as “the”, “of”, “to”
- are typically removed from analysis

https://www.tidytextmining.com

When doing sentiment analysis…

stemming - lexicon normalization

- Identifying the root for each token
- Jumping, jumped, jumps, jump all have the same root ‘jump’
- Where things get tricky: jumper???

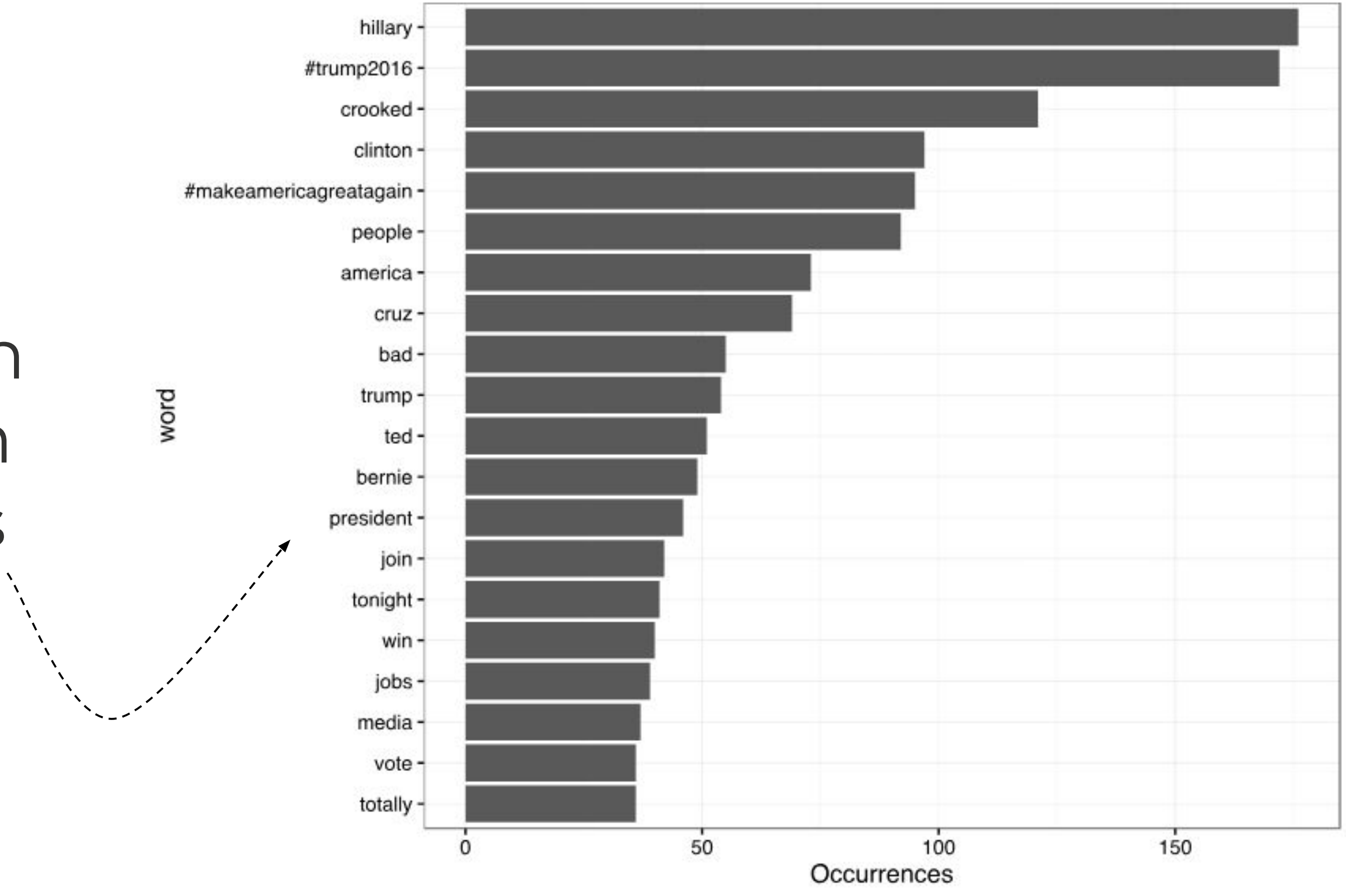
In text analysis, your choices matter:

1. How to tokenize?
2. What lexicon to use?
3. Remove stop words? Remove common words?
4. Use stemming?

Are the angrier and more hyperbolic tweets from Trump himself (rather than his staff)?

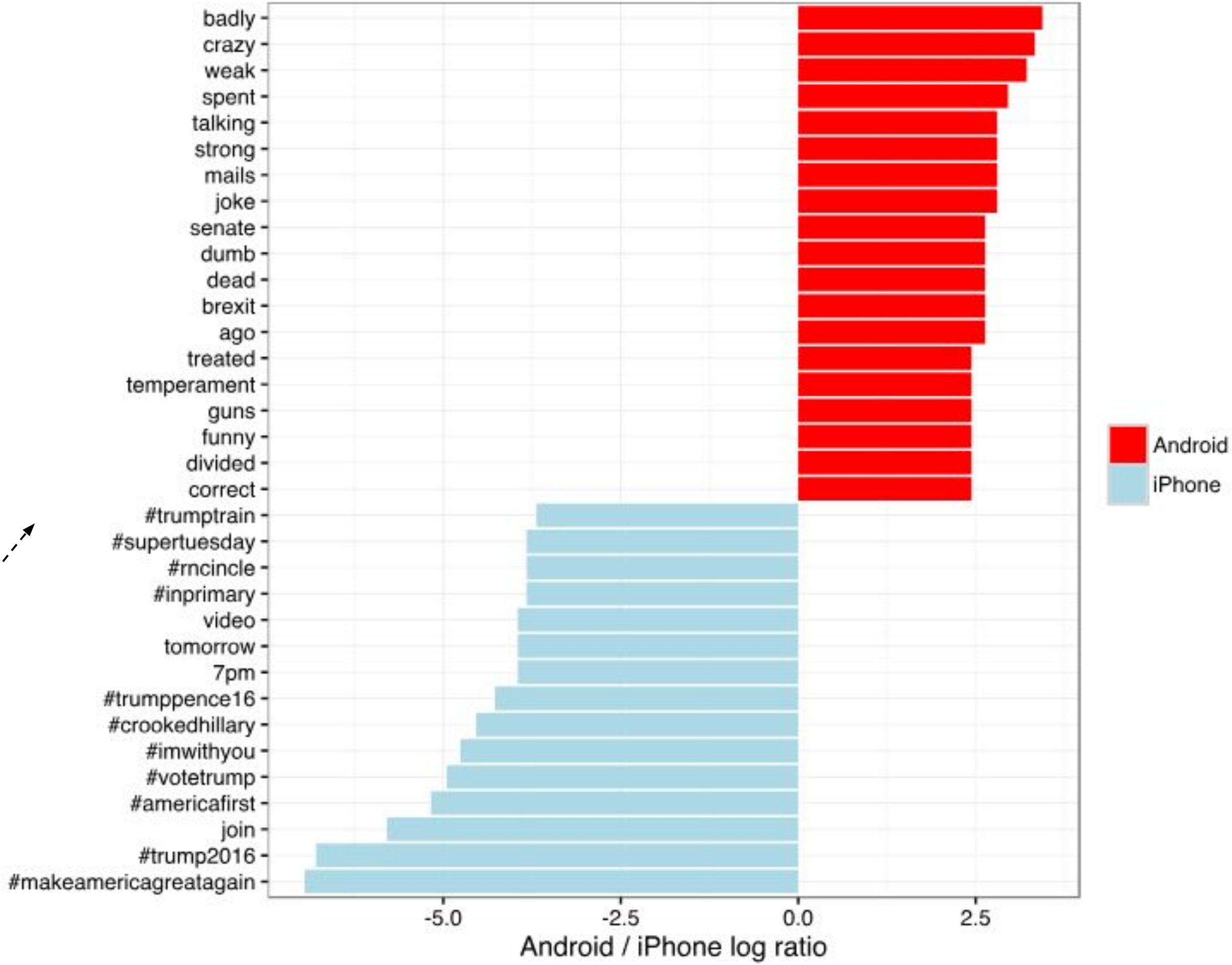
(Note: we knew Trump was using a Samsung Galaxy)

Most common words in Trump’s tweets

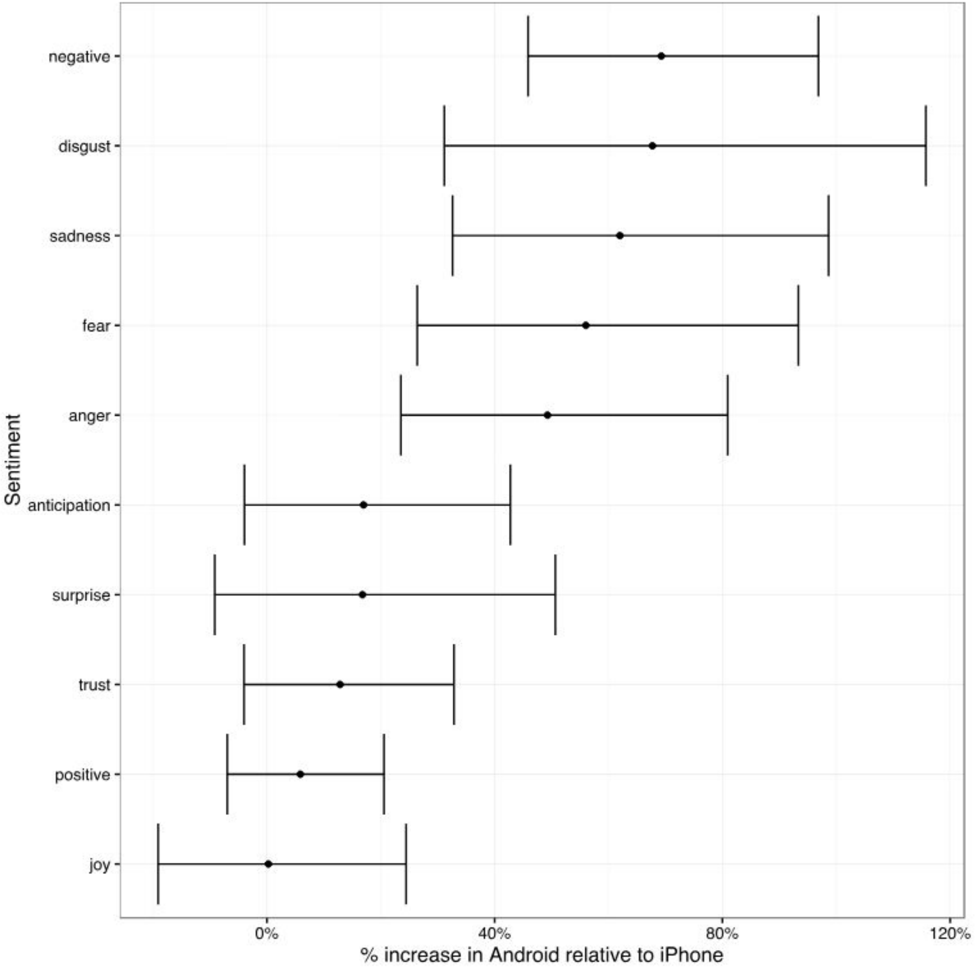


http://varianceexplained.org/r/trump-tweets/

Frequency broken down by device



http://varianceexplained.org/r/trump-tweets/

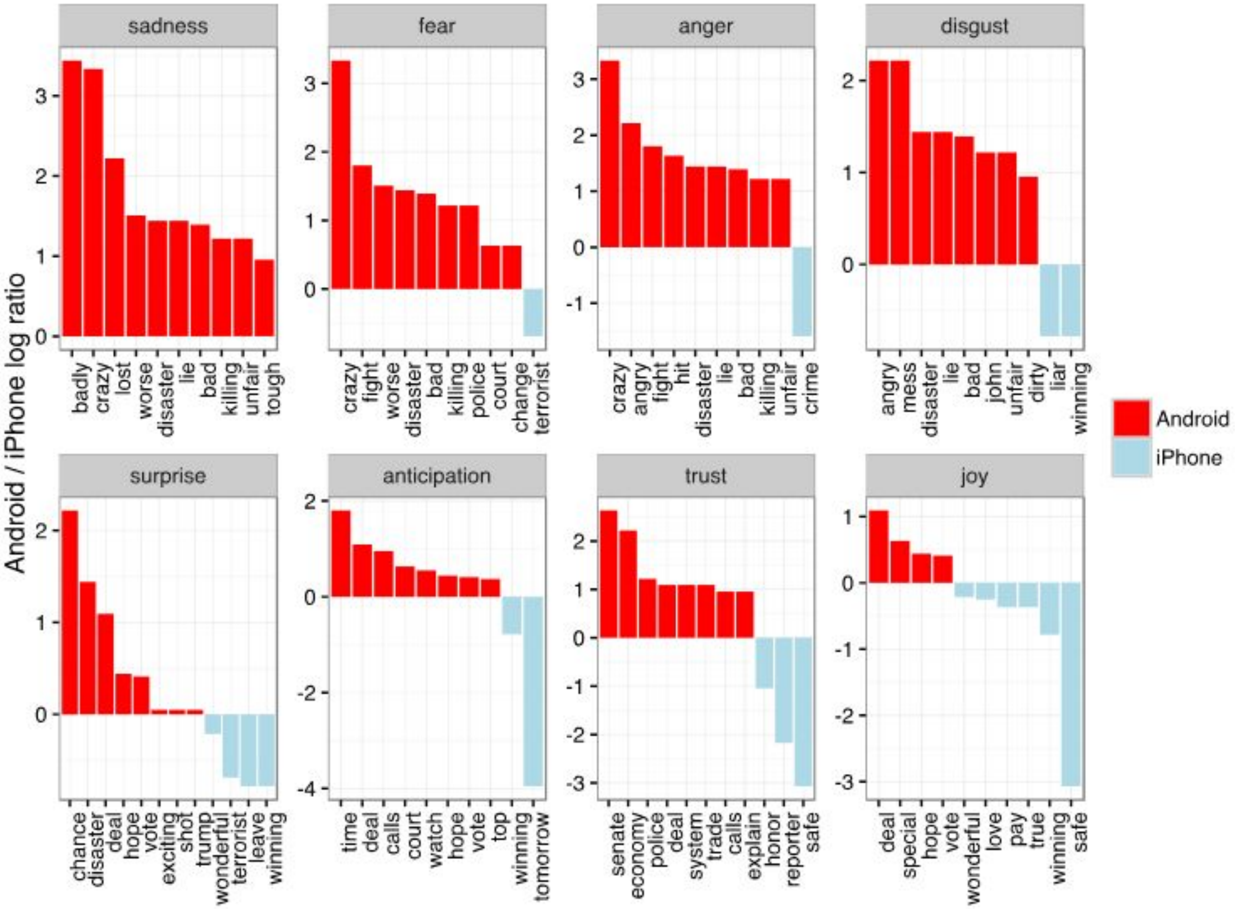


Emotionally charged negative sentiment words more frequently sent from an android

"Trump's Android account uses about 40-80% more words related to disgust, sadness, fear, anger, and other "negative" sentiments than the iPhone account does"

http://varianceexplained.org/r/trump-tweets/

Display of words driving this increase in negative sentiment

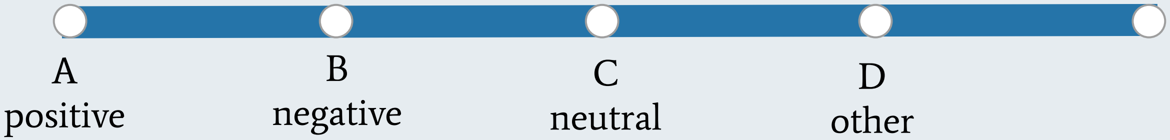


http://varianceexplained.org/r/trump-tweets/

Sentiment Limitations

How would you classify the sentiment of the following sentence?

“The idea behind the movie was great, but it could have been better”



A
positive

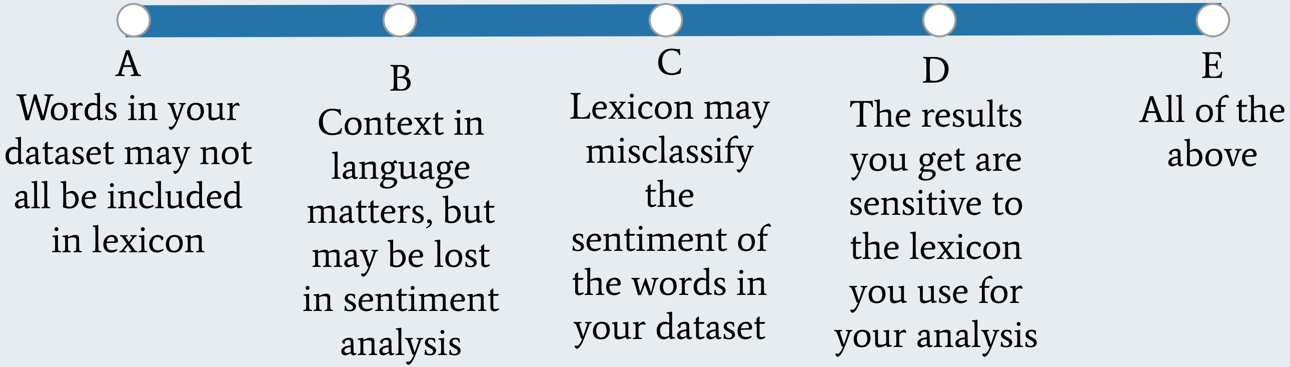
B
negative

C
neutral

D
other

Sentiment Limitations

What is a limitation of sentiment analysis?



A
Words in your dataset may not all be included in lexicon

B
Context in language matters, but may be lost in sentiment analysis

C
Lexicon may misclassify the sentiment of the words in your dataset

D
The results you get are sensitive to the lexicon you use for your analysis

E
All of the above