Demonstration of automated Natural Language Processing

Ian Hussey[*]

14 November 2017

This is a very simple demonstation of the automatic classification of text according to predefined word categories. This is sometimes referred to as Natural Language Processing, Sentiment Analysis (when the valence of words is be analysed), or a few other terms.

I purposefully chose a data source that was too large to analyse by hand: Data was scraped from a Reddit.com/r/askreddit thread on individuals' first thought upon waking up after surviving a suicide attempt. c.2000 responses with a mean of c.120 words each.

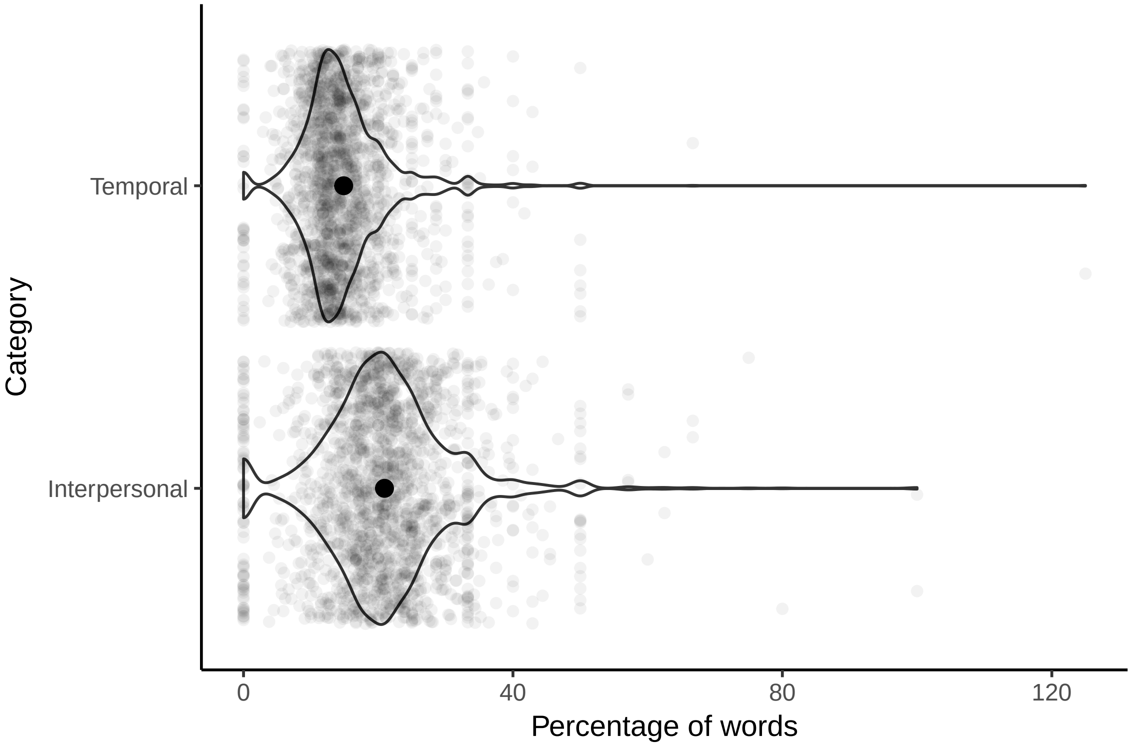
Categorisations are done at the "parsel" level, where a parself represents an individual comment. A parsel could equally represent questions and answers in an exchange, utterances within set time periods, or therapeutic sessions.

The below simply plots and statistically compares the proportion of words that are made up by different categories. Many other, more complex analyses are also possilbe. E.g., network analyses of specific words, or categories (e.g., what words cooccured with what others in a given statement), analyses of the temporal dynamics of categories across time, or between groups, etc.

Deeper semantic understanding is also possible. E.g., the below counts "not bad" as an instance of a negative word (bad) rather than a positive meaning (good). Other libraries can pull out this deeper meaning among words. The below serves as an accessible demo however.

[*] Ghent University. Email: ian.hussey@ugent.be

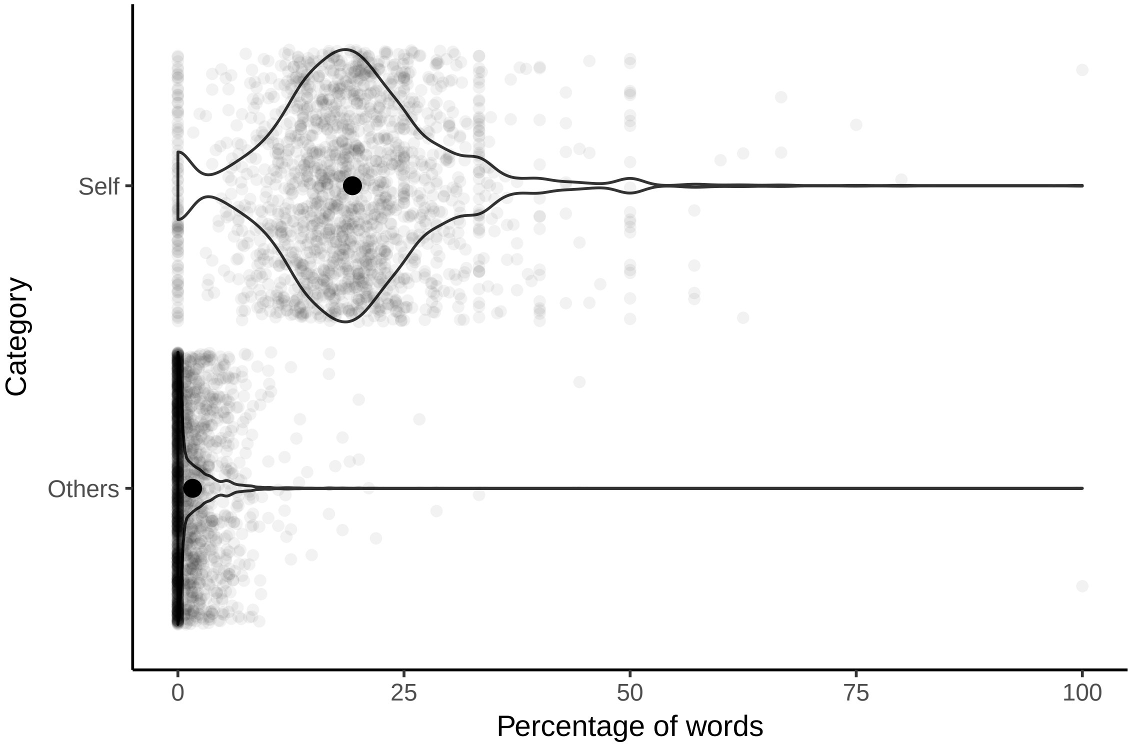
Interpersonal vs temporal



```
##
##  Welch Two Sample t-test
##
## data:  percent by category
## t = 21.219, df = 3228.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.502421 6.622818
## sample estimates:
## mean in group Interpersonal      mean in group Temporal
##                    20.92940                    14.86678
```

```
##
## Cohen's d
##
## d estimate: 0.7114699 (medium)
## 95 percent confidence interval:
##       inf       sup
## 0.6436830 0.7792567
```

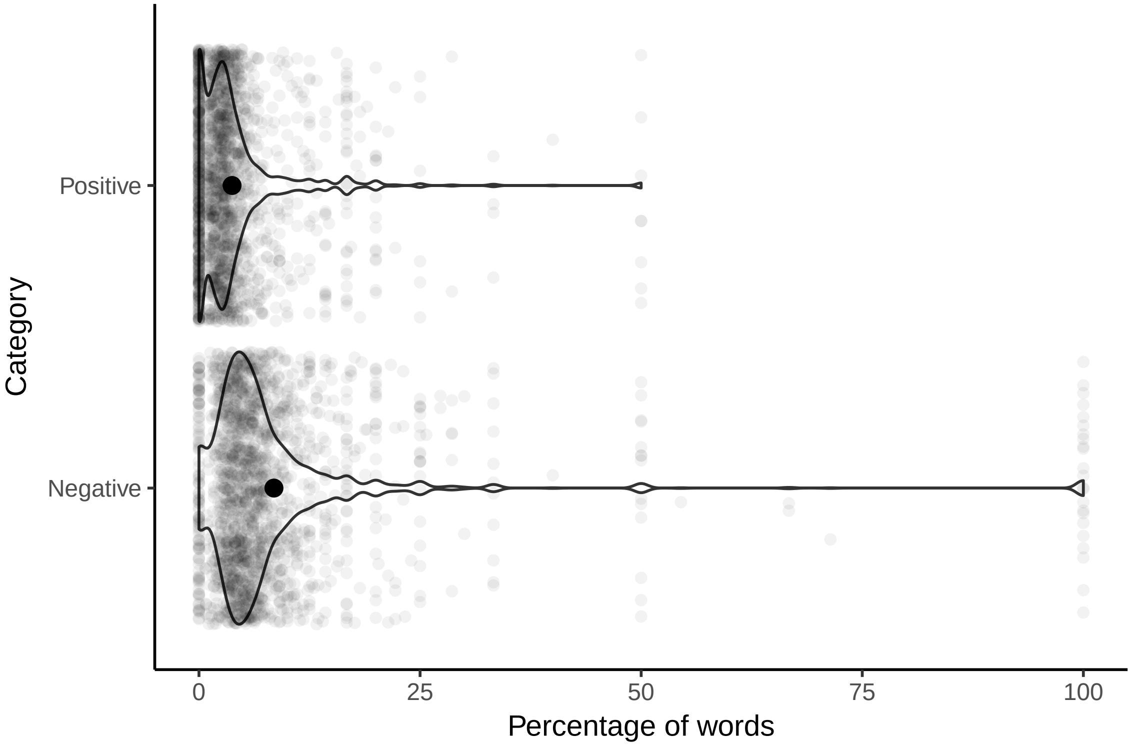
Interpersonal self vs others



```
##
##  Welch Two Sample t-test
##
## data:  percent by category
## t = -70.651, df = 2309.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.16078 -17.17986
## sample estimates:
## mean in group Others   mean in group Self
##             1.629399            19.299719
```

```
##
## Cohen's d
##
## d estimate: 2.368883 (large)
## 95 percent confidence interval:
##      inf      sup
## 2.283133 2.454633
```

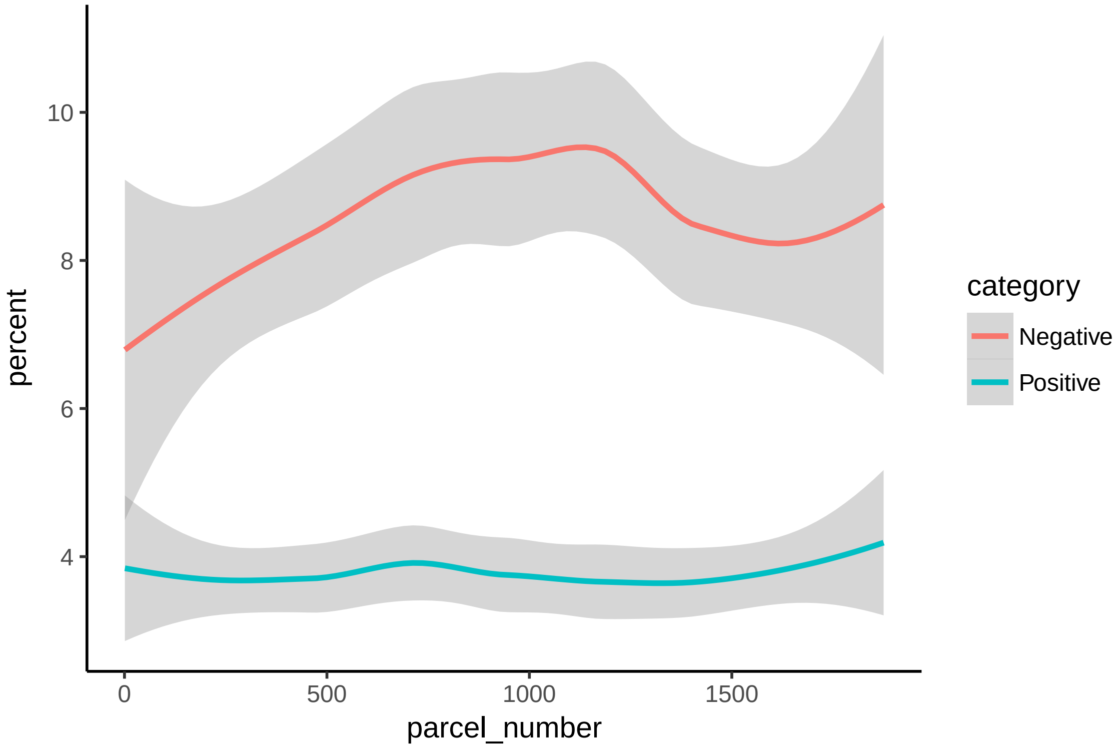
Positive vs negative valence/sentiment



```
##
##  Welch Two Sample t-test
##
## data:  percent by category
## t = 14.126, df = 2349.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4.075130 5.388888
## sample estimates:
## mean in group Negative mean in group Positive
##               8.477835               3.745826
```

```
##
## Cohen's d
##
## d estimate: 0.4793437 (small)
## 95 percent confidence interval:
##       inf       sup
## 0.4118656 0.5468219
```

More complex plots and analyses are of course possible. E.g., changes across parcels (i.e., across time, within a conversation, session, etc)


10
8
6
4
percent
0
500
1000
1500
parcel_number
category
Negative
Positive