Text Mining: Week 5

Joris van Vugt, s4279859

September 29, 2016

1 Exercise 1: Inter-annotator agreement

1.1 Introduction

The goal of this exercise is to do sentiment analysis on comments from a reddit thread. Those sentiments are then compared to those of another annotator using an agreement table and the Cohen's κ measure.

1.2 Difficulties

The comments from the reddit thread are sometimes hard to annotate, even for humans. For example, a few of the posts were deleted. I will give a short list of some of the other difficulties:

- **Reactions to other posts**
  The content of the parent message sometimes has to be known to figure out the sentiment of the reaction. For example, '*Have you seen it?*' in reply to a positive comment, would imply negative feelings with the author.
- **Inside jokes**
  Comments with inside jokes (e.g., '*10/10 comment m8*') also pose problems for sentiment analysis. Reddit is full of these 'memes' which require non-standard interpretations of text.
- **World knowledge**
  Comments that require background knowledge about other concepts will also be hard to classify. For example, '*Who the fuck cares? It's Teenage Mutant Ninja Turtles.*'. Apart from the bad language, this comment hardly seems to convey any sentiment, on the surface. One would have to know that most people really liked TMNT in their childhoods and infer that this person would like to see another film.

1.3 Agreement table and Cohen's κ

My fellow data scientist Tanja Crijns was kind enough to give me her annotations of the comments. Comparing these to my own annotations yields the agreement table shown in Table 1

$$\kappa = \frac{p(a) - p(e)}{1 - p(e)}$$

Agreement table		Tanja	
		P	N
Joris	P	19	10
	N	8	13

Table 1: Agreement table

Using the values from the agreement table:

$$
\begin{align}
p(a) &= \frac{19+13}{19+8+10+13} = \frac{32}{50} = 0.64 \tag{1}\\
p(e, \text{yes} \mid \text{Joris}) &= \frac{19+10}{8+13+19+10} = \frac{29}{50} = 0.58 \tag{2}\\
p(e, \text{yes} \mid \text{Tanja}) &= \frac{19+8}{10+13+19+8} = \frac{27}{50} = 0.54 \tag{3}\\
p(e, \text{yes}) &= 0.58 \times 0.54 = 0.3132 \tag{4}\\
p(e, \text{no}) &= 0.42 \times 0.46 = 0.1932 \tag{5}\\
p(e) &= 0.3132 + 0.1932 = 0.5064 \tag{6}\\
\kappa &= \frac{0.64 - 0.5064}{1 - 0.5064} \approx \frac{0.1336}{0.4936} \approx 0.27 \tag{7}
\end{align}
$$

2 Exercise 2: Classifier evaluation

a. i. Precision(pos) $= \frac{10}{10+4} = \frac{10}{14} \approx 0.71$

   ii. Recall(pos) $= \frac{10}{10+14} = \frac{10}{24} \approx 0.42$

b. i. Precision(neg) $= \frac{22}{22+14} = \frac{22}{36} \approx 0.61$

   ii. Recall(neg) $= \frac{22}{22+4} = \frac{22}{26} \approx 0.85$

c. i. Macro-Precision $\approx \frac{0.71+0.61}{2} \approx 0.66$

   ii. Macro-Recall $\approx \frac{0.42+0.85}{2} \approx 0.64$

d. i. Micro-Precision $= \frac{10+22}{10+4+14+22} = \frac{32}{50} = 0.64$

   ii. Micro-Recall $= \frac{10+22}{10+4+14+22} = \frac{32}{50} = 0.64$