# Cross-discourse development of supervised sentiment analysis

Phil Smith
pxs697@cs.bham.ac.uk

Supervisor: Dr. Mark Lee

February 17, 2012

# What is Sentiment Analysis?

- Sentiment Analysis is the computational classification of qualitative and emotional language that is expressed through written discourse.
- Detail in classification can vary:
  - Polarity based classification - Positive or Negative
  - Finer-grained combinations of emotional states: Joy, anger, surprise, fear, sadness..
- Occasionally referred to as Opinion Mining - focuses on product reviews
- Applies knowledge from a number of fields:
  - Natural Language Processing
  - Corpus Linguistics
  - Information Retrieval

# Why is Sentiment Analysis important?

*"Emotion is one of the most central and pervasive aspects of human experience"*
*Ortony, Clore & Collins (1988)*

- Decisions are not always motivated by logic alone, and can often be overpowered by emotion.
- The consequences of not understanding or misunderstanding emotion can be costly.

# Practical Applications

- Customer-facing industries
- Advertising campaigns
- Augmenting search capabilities
- Personal decision making

# Where can it be applied?

- Social media: Facebook, Twitter, blogs, forums, reviews.
- Customer Feedback: Call centres, emails, text messages, web surveys.
- Traditional Media: News headlines and articles

# Examples from SemEval 2007 Headline Corpus

*"Bombers kill shoppers"*

- Clearly negative.

*"PS3 pricey but impressive"*

- Comparative statement, weight put on final word.
- Lexical order is important to understanding the sentiment here.

*"Nigeria hostage feared dead is freed "*

- Again, interpretation depends on final word.
- The majority of the statement suggests a negative statement.

# Sentiment Analysis Corpora

- Pang & Lee (2002) Movie Review Corpora ($\sim$2,000).
- Semeval 2013 SMS and Tweet Corpora ($\sim$12,000).
- Lin and Hauptmann (2006) Political Perspective Corpora.
- CMC Suicide Note Corpora.

# Data Dot Gov

- Making Public Data Public
- ~5,400 Datasets available from central government departments.
  - Department for Transport.
  - Department for Communities and Local Government.
  - Office for National Statistics.
  - Department of Health.
- Make clear results of policy changes.
- Can be used to analyse trends over time.

# NHS Choices Patient Feedback

- Public given the opportunity to submit feedback online regarding the NHS.
- Post about experiences with individual hospitals and GP surgeries.
- Responses from the relevant organizations are also able to be given.
- Currently there are 3 main datasets:
  1. Overall hospital ratings.
  2. Individual comments about hospitals and responses.
  3. Individual comments about GPs and responses.

# Individual Comments About Hospitals and Responses

- Two sub-datasets within this:
  1. Feedback from the period $1^{st}$ January 2008 to $31^{st}$ July 2011.
  2. Feedback from $1^{st}$ August to today.
- Available as comma-separated value files.
- The most recent dataset was chosen to derive corpora for preliminary experiments.

# Structure of the Dataset

- Hospital Name - ∼450 Unique Hospitals and Treatment Centres.
- Relative trust and organisation codes.
- Comment ID
- Comment Title
- Liked
- Disliked
- Advice
- Date of comment
- Date of patient admission
- Response
- Date of Response

# Creating the Corpus

- Extract relevant data for preliminary experimentation.
- Liked, Disliked and Advice Data

| Corpus | # of Docs | # of words | $Doc_{avglength}$ | # of unique words |
|---|---|---|---|---|
| *Expressive* | | | | |
| Positive | 1152 | 75052 | 65.15 | 6107 |
| Negative | 1108 | 76062 | 68.65 | 6791 |
| *Persuasive* | | | | |
| Positive | 768 | 46642 | 60.73 | 4679 |
| Negative | 864 | 113632 | 131.52 | 7943 |

- NB. Advice data was manually annotated.

# Annotating the Persuasive Corpus

- 1702 unique documents were annotated.
- Annotations were placed at the document level.
- Overall sentiment of the recommendation taken into account.
  - In the cases of comparative phrases being used to express sentimental information, the predominant sentiment was chosen.
- 70 were annotated as not containing any emotion due to the content.
  - e.g "N/A"
- Neutrally annotated data was removed from the corpus.

# Examples from the Corpus

*Liked:* *"The honesty of the Consultants PA's."*

*Disliked:* *"The attitude of the nursing staff."*

*Positive Advice:* *"Please just do what you do, you are all so good."*

*Negative Advice:* *"It's very well spending all this money on a new hospital building, however what is the point when people are being wrongly diagnosed."*

# Keyword List: Positively Annotated Sections of Corporus: Unusually Frequent

Comparing the Positive section of the Persuasive Corpus with the Positive section of the Expressive Corpus (reference corpus)

| Rank | Keyword | Keyness | Frequency |
|---|---|---|---|
| 1 | thank | 46.631 | 190 |
| 2 | tameside | 27.496 | 14 |
| 3 | scan | 19.602 | 42 |
| 4 | week | 18.520 | 41 |
| 5 | musgrove | 17.676 | 9 |
| 6 | wife | 16.411 | 41 |
| 7 | hospital | 16.102 | 387 |
| 8 | thanks | 14.584 | 72 |
| 9 | rotheram | 13.876 | 10 |
| 10 | cannock | 13.748 | 7 |

# Keyword List: Positively Annotated Sections of Corpora: Unusually Infrequent

| Rank | Keyword | -Keyness | Frequency |
|---|---|---|---|
| 1 | friendly | 44.210 | 57 |
| 2 | professional | 26.628 | 40 |
| 3 | polite | 24.586 | 9 |
| 4 | environment | 16.675 | 5 |
| 5 | helpful | 14.116 | 59 |
| 6 | staff | 13.723 | 466 |
| 7 | including | 13.200 | 3 |
| 8 | caring | 12.912 | 51 |
| 9 | treated | 12.610 | 70 |
| 10 | exceptional | 12.400 | 6 |

# Keyword List: Negatively Annotated Sections of Corpus: Unusually Frequent

| Rank | Keyword | Keyness | Frequency |
|---|---|---|---|
| 1 | hospital | 54.296 | 659 |
| 2 | now | 24.210 | 165 |
| 3 | hip | 23.245 | 37 |
| 4 | catheter | 22.266 | 32 |
| 5 | wrightington | 19.084 | 18 |
| 6 | days | 18.406 | 162 |
| 7 | weeks | 17.933 | 157 |
| 8 | daughter | 17.447 | 66 |
| 9 | doctor | 16.803 | 307 |
| 10 | admitted | 16.233 | 95 |

# Keyword List: Negatively Annotated Sections of Corpus: Unusually Infrequent

| Rank | Keyword | -Keyness | Frequency |
|---|---|---|---|
| 1 | patients | 56.453 | 232 |
| 2 | improved | 50.314 | 18 |
| 3 | staff | 46.085 | 444 |
| 4 | food | 39.336 | 52 |
| 5 | communication | 37.649 | 32 |
| 6 | room | 34.476 | 97 |
| 7 | improvement | 28.515 | 3 |
| 8 | waiting | 25.936 | 208 |
| 9 | little | 19.653 | 54 |
| 10 | training | 17.706 | 10 |

# Concordance Data: Experience

| Rank | Keyword | -Keyness | Frequency |
|---|---|---|---|
| 24 | experience | 11.388 | 107 |

- Positive expressive corpus.

```
l Helpful in my experience. Very nice nursing
this a pleasant experience! nice and relaxed t
 staff made the experience bearable and the bi
j other peoples Experience with midwives in ot
nd she made the experience very happy for me a
uld to make the experience go as smoothly as i
 Throughout the experience all the staff I enc
 happy with the experience. I would recommend
l they made the experience easy to cope with.
```

# Concordance Data: Experience

- Positive persuasive corpus.

```
 me.  My recent experience gave me a very positive
ing the overall experience a little negative.  The
 for making this experience as easy as it could be.
are my positive experience of excellent care recei
ly say that our experience as a family of the care
   A truly great experience! The LGI is a superb ho
 that the whole experience had been very positive.
.      Very good experience. Very welcoming departm
```

# Concordance Data: Negative Experience

| Rank | Keyword | -Keyness | Frequency |
|---|---|---|---|
| 22 | experience | 13.558 | 131 |

- Negative expressive corpus.

```
rembling at this experience and I have nowh
n it.  The whole experience is proving a ni
e improved in my experience at this hospita
d that the whole experience was a waste of
rd. We know from experience that we can exp
pinion--my whole experience can only be des
a terrible birth experience which left me w
 Based on my own experience I cannot think
 based on my own experience.  However, one
```

# Concordance Data: Negative Experience

Negative Persuasive corpus.

```
osted about my experience trying to book a
eeks ago of my experience and still havent
 have had a bad experience today in that I
bout the whole experience.  After being ad
 patient.   My experience in 2010 and sinc
ch, this whole experience totally knocked
y acerbated my experience. I can't speak f
ves Nurses, my experience was so traumatis
e waits, in my experience some poor staff
```

# Applying Supervised Machine Learning Methods

- Treats sentiment analysis as a text classification problem.
- Three machine learning methods were used:
  - Naive Bayes
  - Multinomial Naive Bayes
  - Support Vector Machines
- Various features were used:
  - Unigrams
  - Bigrams
  - Bigrams + Part of Speech information

# Validation of training models

- 10-fold cross-validation performed to check model accuracy:

| Features | NB | Multinomial NB | SVM |
|---|---|---|---|
| Unigrams | **79.65** | **78.14** | **76.11** |
| Bigrams | 57.79 | 60.84 | 63.36 |
| Bigrams + POS | 74.25 | 75.71 | 72.83 |

# Testing on Persuasive Section

- Testing produces the following accuracy results:

| Features | NB | Multinomial NB | SVM |
|---|---|---|---|
| Unigrams | 76.60 | **83.39** | 68.12 |
| Bigrams | 50.10 | 62.13 | 57.6 |
| Bigrams + POS | 66.73 | 70.42 | 64.00 |

# Shortcomings of Machine Learning Methods

- A statistical approach to lanugage
- Does not observe nuances of a language
  - Negation is ignored.
  - Emphasis is ignored.
- Repetition often doesn't imply a particular sentiment.

# Discussion

- It would appear that training on a corpus whose discourse function differs to that of the training corpus still enables machine learning algorithms to create suitable models.
- Word frequency should be taken into account when training the models.
- Multinomial Naive Bayes is a machine learning technique well suited to this.

# Future Work

- Current Corpus: Experiment with other machine learning approaches and features.
- Annotate the larger dataset: $\sim$10,000 rows of data which could be used as the basis for further experimentation.
- Annotate the corpus with fine-grained emotions
- Investigate and present results from suicide note corpus.

# Summary

- Sentiment Analysis
- NHS Choices
- Creating the Expressive and Persuasive Corpus
- Analysing the Corpus
- Training and testing supervised machine learning methods