Predicting Outcomes from Language in Mental Health Therapy

Matthew Purver, Queen Mary University of London
Rosemarie McCabe, University of Exeter
Christine Howes, University of Gothenburg

Queen Mary University of London
CIS centre for intelligent sensing
Cognitive Science Research Group
http://cogsci.eecs.qmul.ac.uk

Acknowledgements

The PPAT & AOTD projects were supported by Queen Mary University of London's EPSRC-funded Pump-Priming and Innovation Funds, and PsychologyOnline Ltd; and completed in collaboration with PsychologyOnline Ltd and iLexIR Ltd.

EPSRC
Pioneering research and skills

PsychologyOnline

Queen Mary
University of London

iLexIR
NLP Consultancy

Motivation

- Communication is important in mental health:
  - Communication quality associated with outcomes
    - (Ong et al, 1995; McCabe et al, 2013)
  - Communication *during treatment:*
    - Conversation structure (how)
    - Conversation content (what)

- Natural Language Processing (NLP)
  - techniques for detecting structure and content
  - can NLP techniques help us analyse & understand mental health therapy?

Motivation (2)

- Online text-based therapy recently introduced
- Approved & available via NHS
  - Depression & anxiety
  - Convenient, anonymous
  - As effective as treatment as normal
    - (Kessler et al, 2009)

- Can NLP techniques help us analyse & evaluate this new medium?
  - (especially because it gives us easier access to data)

Study 1: Face-to-Face Dialogue

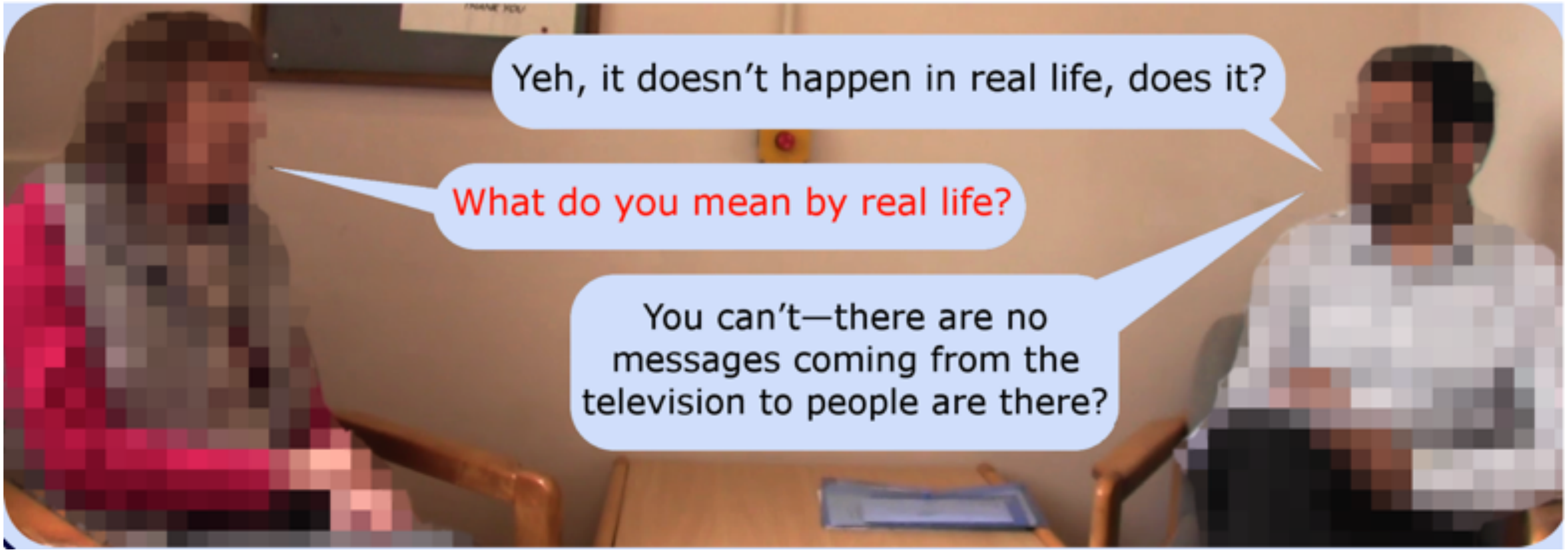
- Transcripts of therapy for schizophrenia
- Symptom measures
  - *positive* (delusions, hallucinations, beliefs)
  - *negative* (withdrawal, blunted affect, alogia)
- Outcome measures
  - ratings of communication quality
  - future adherence to treatment:
    - non-adherence: risk of relapse 3.7 times higher
- Manual annotation & statistical analysis
  - McCabe et al (2013)
- Automatic NLP processing & machine learning
  - Howes et al (2012; 2013)

Content: Topic

- Manual definition of 20 “topics”:
  – Medication, side-effects, treatment, management
  – Symptoms, health, self-harm
  – Daily activities, living situation, relationships, …

Topic	Name	Description
01	Medication	Any discussion of medication, excluding side effects
02	Medication side effects	Side effects of medication
03	Daily activities	Includes activities such as education, employment, h
04	Living situation	The life situation of the patient, including housing, fi
05	Psychotic symptoms	Discussion on symptoms of psychosis such as halluc
06	Physical health	Any discussion on general physical health, physical i
07	Non-psychotic symptoms	Discussion of mood symptoms, anxiety, obsessions,
08	Suicide and self harm	Intent, attempts or thoughts of self harm or suicide (p
09	Alcohol, drugs & smoking	Current or past use of alcohol, drugs or cigarettes an
10	Past illness	Discussion of past history of psychiatric illnesses, in
11	Mental health services	Care coordinator, community psychiatric nurse, soci
12	Other services	Primary care services, social services, DVLA, emplo

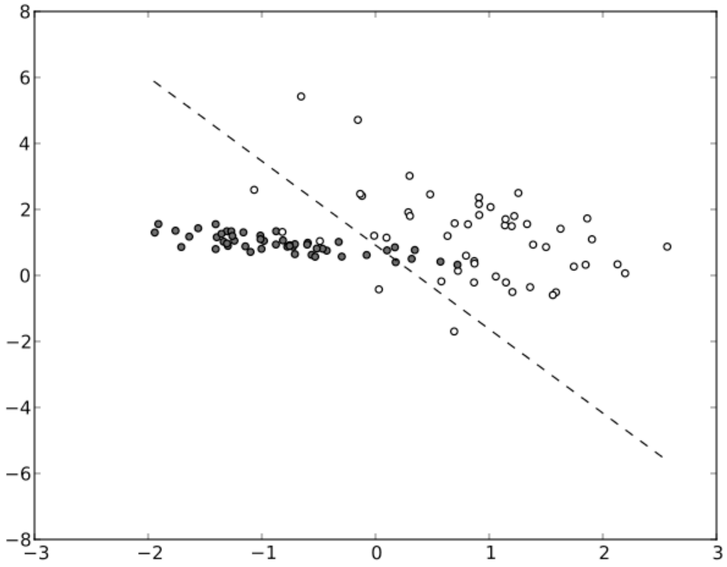
Structure: Repair



- Manual linguistic analysis
  - Significant role of *repair*
  - Patient-initiated other-repair & self-repair – but rare (0.8%)
- What else might be important? (content / structure?)

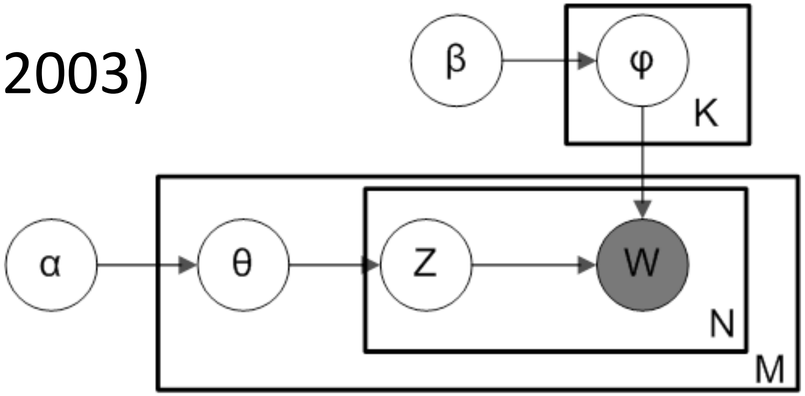
N-gram models

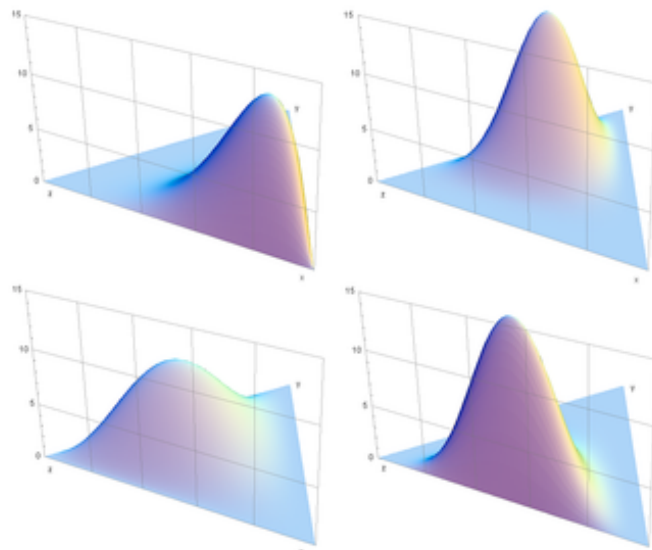
- Extract all 1-, 2-, 3-, ... word sequences
  - (with frequency cut-off)
  - approximation of content
  - approximation of structure
- Very high-dimensional feature space
  - Learn correlations automatically
  - A “brute force” approach



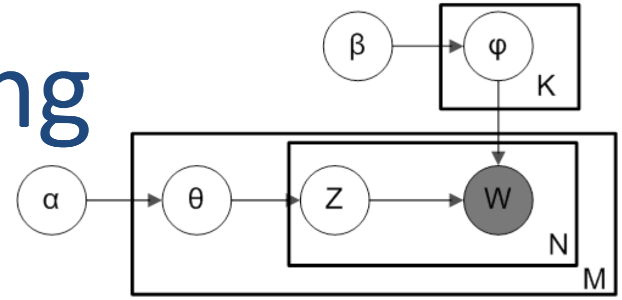
Topic Modelling

- Latent Dirichlet Allocation (Blei et al, 2003)
- Unsupervised Bayesian model:
  - texts as mixtures of “topics”
  - topics as distributions over words
- No prior knowledge of topics
  - number of topics
  - likely distribution shapes
  - (automatically optimised)
- Successful application in a wide range of domains & tasks





Automatic topic modelling



- Infer 20 lexical “topics”:

Topic 0	feel low alright mood long drug feeling tired time confide
Topic 4	voices pills mood cannabis telly voice shaking chris contro
Topic 5	letter health advice letters council copy send dla cpn prob
Topic 7	church voice voices hear medication sister bad hearing tak
Topic 9	school children kids back september oclock gonna phone
Topic 10	weight months medication stone risk lose eat write gp has
Topic 11	place support work centre gotta job stress feel psychologis
Topic 12	door house police thought ring knew worse wall hadnt sat
Topic 13	doctor alright years nice ill anxious write long sit eye hear
Topic 14	drug taking milligrams hundred doctor night time medicat
Topic 15	sort medication work drugs kind team issues drink alcohol
Topic 16	mum place brother tablets died dad depot house meet mo
Topic 17	people life drug make care lot friends dry camera live cop
Topic 18	alright house drink drinking money alcohol god drugs livin

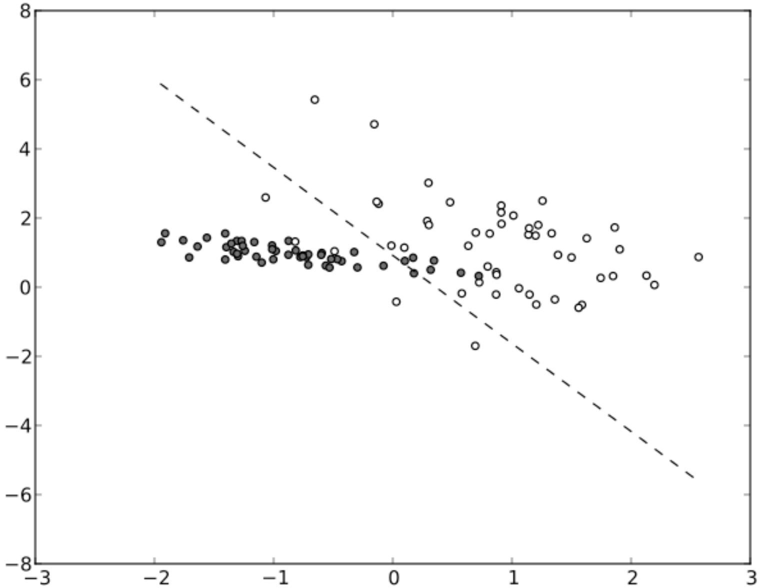
Topic modelling

- Compare to manually defined topics
- LDA topics given manual “interpretations” & compared:
  - (including sentiment aspect)

	Interpretation	Example words from top 2
0	Sectioning/crisis	hospital, police, locked
1	Physical health - side-effects of medication and other	gp, injection, operation
2	Non-medical services - liaising with other services	letter, dla, housing
3	Ranting - negative descriptions of lifestyle etc	bloody, cope, mental
4	Meaningful activities - social functioning	progress, work, friends
5	Making sense of psychosis	god, talking, reason
6	Sleep patterns	sleep, bed, night
7	Social stressors - other people stressors/helpful	home, thought, told
8	Physical symptoms - e.g. pain, hyperventilating	breathing, breathe, burning
9	Physical tests - Anxiety/stress arising from tests	blood, tests, stress
10	Psychotic symptoms - e.g. voices, etc.	voices, hearing, evil
11	Reasurrance/positive feedback/progress	sort, work, sense
12	Substance use - alcohol/drugs	drinking, alcohol, cannabis

Prediction Experiments

- Standard supervised machine learning via Weka (Hall et al, 2009)
  - Decision trees
    - Qualitative inspection
  - Logistic regression
    - Better fit (usually)
  - Support vector machines
    - High-dimensional feature spaces



- Predict outcomes:
  - Measure accuracy as F-score
    - Recall = true positive rate (sensitivity)
    - Precision = positive predictive value)
    - Weighted average over classes
- 10-fold cross-validation
  - Train on 90%, test on 10%
  - Repeat x10 to cover all data

Results – non-adherence prediction

- Classification experiments (SVMs, 10-fold cross-validation)
- Predicting over entire dialogues (patient turns only):

Features	P (%)	R (%)	F (%)
Class of interest	28.9	100.0	**44.8**
Best ngram features	70.3	70.3	**70.3**

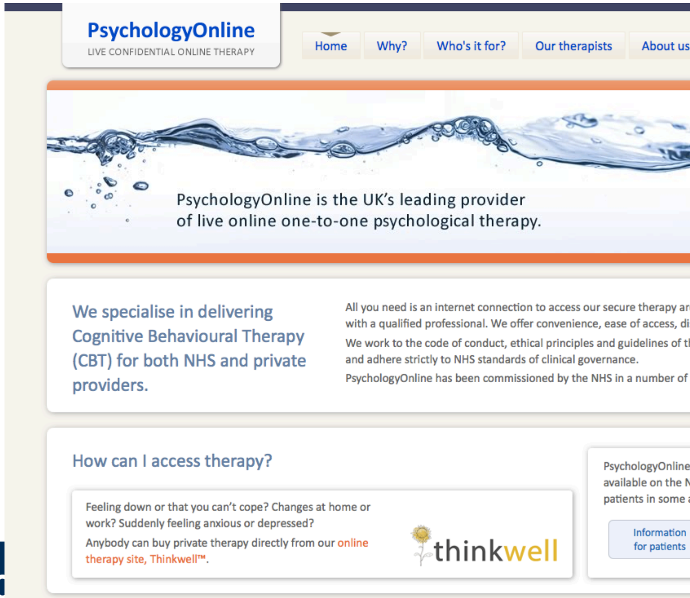
- Accuracy for automatic topics 54.1%; manual topics 66.2%
- (Similar for symptoms, some outcomes e.g. HAS, PEQ)
- Human psychiatrist given same task:

Data	P (%)	R (%)	F (%)
Text transcripts	60.3	79.6	**68.6**
Transcripts + video	69.6	88.6	**78.0**

- But how well will this generalise?

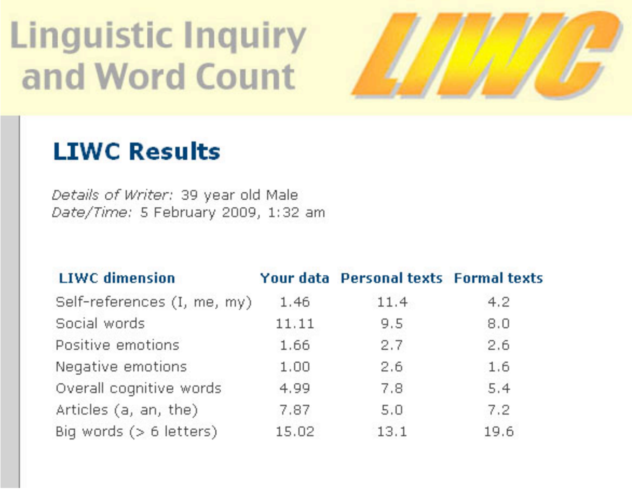
Study 2: Online Therapy

- Online, anonymous, text-based therapy for depression and anxiety (PsychologyOnline Ltd)
  - Cognitive Behavioural Therapy (CBT)
  - 500 patients (352 female, 146 male, 2 unknown), 64 therapists
  - 2066 sessions, 1864 from ongoing or complete treatment
  - mean 5.65 sessions per patient (min 1, max 15)
- Anonymisation independently (via iLexIR Ltd)
  - Using RASP toolkit (Briscoe et al, 2006)
    - Person & organisation names, places, dates
    - Harder than standard text tasks …
    - … so some errors, manually corrected
- Outcome measures
  - Patient Health Questionnaire (PHQ-9)
  - 0-27 scale: moderate/severe >=10
    - (in/out-of-caseness)
  - Progress: change since start



Sentiment/Emotion Detection

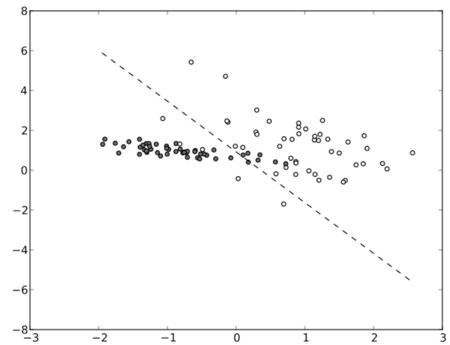
- Detect emotional content
  - positive & negative sentiment
  - anger
  - challenge & emotion elicitation in CBT
- Compared 3 existing tools
  - 1 dictionary-based: LIWC
  - 2 data-based: Stanford (news), Sentimental (social media)
- Accuracy vs manual sentiment annotations
  - LIWC 34-45%
  - Stanford 51-54%
  - Sentimental 63-80%


Linguistic Inquiry and Word Count LIWC

LIWC Results

Details of Writer: 39 year old Male
Date/Time: 5 February 2009, 1:32 am

LIWC dimension	Your data	Personal texts	Formal texts
Self-references (I, me, my)	1.46	11.4	4.2
Social words	11.11	9.5	8.0
Positive emotions	1.66	2.7	2.6
Negative emotions	1.00	2.6	1.6
Overall cognitive words	4.99	7.8	5.4
Articles (a, an, the)	7.87	5.0	7.2
Big words (> 6 letters)	15.02	13.1	19.6



Topics vs severity & progress

0	Materials, self-help, procedures	−		10	Unhelpful thinking/habits		
1	Feelings/effects of relationships on sense of self	+	+	11	Work/training/education issues/ goals		
2	Positive reactions/encouragement			12	Agenda/goal setting & review		
3	Issues around food			13	Panic attack description/explanation	−	−
4	Family/relationships & issues with (mostly negative)	+		14	Other healthcare professionals, crises, risk, interventions	++	
5	Responses to social situations			15	Sleep/daily routine	+	
6	Breaking things down into steps	+		16	Positive progress, improvements	−−	−
7	Worries/fears/anxieties	−		17	Feelings, specific occasions/thoughts		
8	Managing negative thoughts/ mindfulness			18	Explaining/framing in terms of CBT model		+
9	Fears, checking, rituals, phobias	−	−	19	Techniques for taking control	−	−

Sentiment/Emotion vs PHQ

	Severity (PHQ)	Progress (ΔPHQ)
Sentiment mean	--	-
Sentiment std dev		+
Anger mean/max	+	
Anger std dev	+	

- More positive sentiment ➔ better PHQ, progress
- More variable sentiment ➔ worse progress
- More/more variable anger ➔ worse PHQ

Predicting final outcomes

- Changes in levels help predict final in/out-of-caseness:
  - using features from initial and/or final sessions:

	Final In-caseness
Baseline proportion	*26.8%* *0.11*
First + last session features, incl deltas	**0.71**
Including early PHQ scores	**0.76**

- Most indicative features:
  - Levels of anger, progress & crisis/risk topics
  - PHQ scores at assessment and initial treatment sessions

Predicting dropout

- Can we predict dropout & non-engagement?
  - 148 of 500 did not enter or stay in treatment

	Dropout
Baseline proportion	*29.6%* *0.14*
Assessment session features	0.65
Treatment session features	**0.70**
Both sessions	**0.73**

- >70% accuracy using initial session features
  - needs to include fine-grained word/ngram features

General Summary

- Linguistic features can predict outcomes:
  - adherence, symptoms, progress
  - and we can extract them reasonably well
- Choice of method and representation is important
  - robust machine learning
  - we'd prefer meaningful representations
  - unsupervised methods to discover topic information
- Conversation structure needs investigation
  - provide a more interpretable model
  - understand the role of therapy structure