

Problems and Possibilities of Corpus Linguistics and Sentiment Analysis in the Health Services

Keith Stuart

Universitat Politècnica de València

kstuart@idm.upv.es

1. GOAL OF RESEARCH

- To automate the analysis of patient feedback on health services (analysis of a corpus of patient narratives about experiences/treatment in the NHS)
- By using Corpus Linguistics and Sentiment Analysis techniques, the health services could ensure that evaluative comments were not missed.
- This could make a difference to the running of hospitals and the well-being of patients.
- A project in collaboration with Patient Opinion:
<https://www.patientopinion.org.uk/info/about>

2. PROJECT DESCRIPTION: SENTIMENT ANALYSIS

- Sentiment Analysis tries to measure subjectivity and opinion in text, usually by capturing speaker/writer evaluations (positive, negative or neutral)
- It tries to measure the strength of these evaluations (the degree to which the word, phrase, sentence, or document in question is positive or negative).
- The task of automatically classifying the polarity (whether the expressed opinion is positive or negative) of texts at the word, phrase, sentence, or document level can be a challenging task.

- In particular, there are problems with sentences such as the following:
 - Admission was ***haphazard*** although the staff were ***very nice*** but ***very busy***.
(negative→positive→implicit negative)
 - ***I would have liked more information*** about what I can or shouldn't do once home for the first few days, and ***information*** regarding my follow up appointment ***is rather vague*** with no number to ring if I need assurance as I live alone.
(implicitly negative)

- What makes Sentiment Analysis difficult is how natural languages are used for communicative purposes. Words in natural languages are dynamic contributors to a process of meaning creation which is strongly affected by the context of use. Hence, a word is a dynamic variable whose value may change depending on the context in which it is used.

Invented example:

- The battery lasts a ***long time***. (positive sentiment)
- The film lasts a ***long time***. ??

Patient Opinion Corpus:

- This is my first experience for a ***long time*** with the NHS, & assuming that today's experience is mirrored across the system, we have an NHS staffed by workers we should be proud of. (neutral, followed by positive sentiment)
- The only disadvantage was having to wait a ***long time*** for an available bed in the ward. (negative sentiment)

2. PROJECT DESCRIPTION: COLLECTING THE DATA

- POC Corpus (Patient Opinion Corpus)

Total number of words in POC	Types (distinct words in POC)	Total number of texts in POC	Time Period (diachronic corpus)
7,327,385	57, 963	50,000	Data collected from 2008 to 2014 by Patient Opinion.org

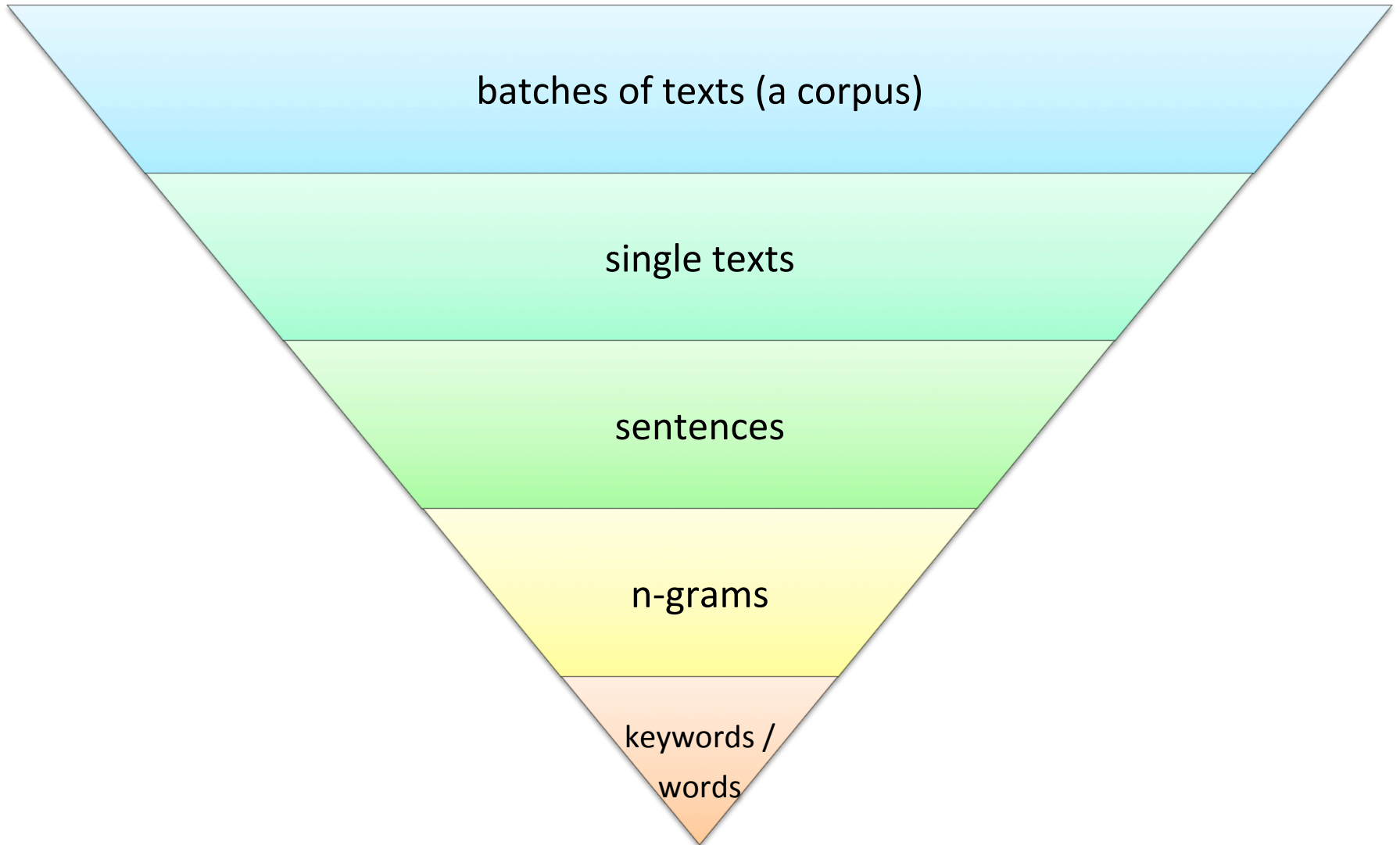
PATIENT OPINION CORPUS DATABASE

- A fairly large Excel file with sentiment scores for 57,963 words was prepared. The manual scoring of weights for the words in the database was carried out using the following criteria:
 - All functional words would have a weight of 0.6
 - Evaluative words would have a weight between + 5 to -5
 - “Neutral” words would have a weight of 1

EVALUATIVE WORDS IN POC

+5	+4	+3	+2	+1	-1	-2	-3	-4	-5
Exemplary Optimal Outstanding Perfect Superb	Excellent Fabulous Fantastic Grateful Lovely	Efficient Friendly Kind Pleasant Professional	Clean Clear Good Nice Polite	Neutral words	Cold Crowded Expensive Tight Tough	Bad Concerned Difficult Late Wrong	Agitated Disrespectful Incompetent Inedible Insensitive	Dire Distraught Furious Undignified Violent	Atrocious Barbaric Excruciating Horrendous Outrageous

3. PROPOSED SOLUTION



3. PROPOSED SOLUTION

- ❑ There(0.6) just(2.0) wasn't(-2.0) enough(2.0) of(0.6) them(0.6) to(0.6) cope(-2.0) with(0.6) patients(1.0) especially(2.0) at(0.6) night(1.0) and(0.6) even(2.0) worse(-3.0) on(0.6) the(0.6) weekends(1.0)
- ❑ Sentence weight: -1.935; Mean: -0.102

4. LINGUISTIC & SEMANTIC PROBLEMS

- Noise
- Semantics
- Idiomatic use of language
- Swear words / Slang
- Intensifiers, Conditionals, Negatives, Comparisons, Modals

//LR 2: Modal perfect rule: would have, may have, might have, should have, could have

- **Noise:**
 - The need to address the problem of noise in text collections is clear as it is detrimental to data analysis. It can be due to aspects such as **typographic and spelling errors** (for example, *actually, actaully, acctually, actualy, acctually*, etc.), informal language, abbreviations among other things. It usually lowers the **data quality** so that the text becomes less accessible to automated processing by a computer.

AVERY = a very
ADAY = a day
AGREAT = a great
AGOOD = a good
ANURSE = a nurse
ABUSY = a busy

ADELAY
AFRACTURE
AFURTHER
APERSON
ASPECIAL
ATICKET

AWEEK
ABAD
ABADLY
ABEAUTIFUL
ACAST
ACAT

ACHEST
ACOMPLETE
ACUBICLE
ACUP
ADIFFICULT
ADR

AFOREST
AGOWN
AHORRIBLE
AJOKE
ALACK
ALARGE

Example of noise in the data

(the indefinite article plus adjective/noun have been typed together)

- Semantics

- //LR3 Reduced distress rule: reduc*, lessen*, decreas*, diminish*, declin*, drop*, stop*
- //LR4 Increased distress rule: grow*, increas*, enlarg*, expand*, buil* up, add*

- continued to sit there for over 2 hours waiting to be called, **my pain gradually increasing** all the time.
- I went back to my husbands bedside but was feeling **increasingly uncomfortable with the situation.**
- I was 41 weeks pregnant exactly and my labour was induced as **I had become increasingly worried** about reduced movements even though looking back on it...

- //LR 1: Negative idiom rule (Idiomatic use of language)

Negative Idioms	Sentiment Score
BURST IN TO TEARS	-5,00
CANNOT STAND	-4,00
CONCERNED WITH	-2,00
DO NOTHING FOR	-2,00
FED UP	-3,00
FIND FAULT	-3,00
FREAKED OUT	-4,00
HAD ENOUGH OF	-3,00
HORROR STORIES	-4,00
IN A MOOD	-3,00
IN A STATE	-4,00
IN TEARS	-4,00
KNOCKED OUT	-4,00
ON EDGE	-2,00
REDUCED TO TEARS	-5,00
SHAKE UP	-4,00
STRESSED OUT	-3,00
TAKE OUT ON	-2,00
TAKEN ABACK	-3,00
TROUBLE MAKER	-3,00
DOOM AND GLOOM	-4,00
FRIGHTENED TO DEATH	-5,00
GRIM LOOKING	-4,00
HIT THE ROOF	-4,00

up with my GP that I was anaemic, the icing on the cake I suppose. As I said, appointment card, and it would be the icing on the cake if the cafe did provide wonderful and efficient people is the icing on the cake. i went in spinal be torture for a hunngry patient! The icing on the cake was the nurse who and was extremely fatigued. The icing on the cake was they threw my professional, genuinely caring, and - the icing on the cake - really nice people. In and everything is suffering. The final icing on the cake was when my GP for giving me the oppurtunity to reply. Icing on the cake has to be free parking and very hungry! Carrot cake minus the icing and with the density to kill a man room to make a phone call!! The real icing on the cake was to be told that believe that this service put the icing on the cake and I believe it left me my ability to go home later that day. Icing on the cake that day? - lots of it. could be sorted then that would be the icing on the cake for me. It seems such or offer one kind word seemed like the icing on the cake to a bad day at the a speedy recovery...that really was the icing on the cake. THANKYOU. I know

An additional benefit to something already good.

Swear words / Slang

- *Alcoholic* = *alkie*. As it is derogatory -2.
- Slang as intensifiers like *blimey*, *cor*, *golly*, *gosh* +2 .
- *Dig*, it is used as a slang term for love, +3.
- *Crog*, slang for 'assault sexually', -5.
- *Fucking* , *effin* -5.
- *Arsed*, *arse*, *bollocked* are all terms with a negative profile -4.
- *Bugger*, *shit* -4.
- *Pissed*, both 'drunk' and 'angry' -2.
- *WTF*, *WTH* -3.
- Disgust exclamations such as *yeugh*, *yuck*, *eurgh*, *argh*, *urgh*, *grrr*, *yikes...* -2.
- Delight exclamations such as *yippee*, *wow*, *whee*, *hurrah*, *hurray* +2.

Intensifiers

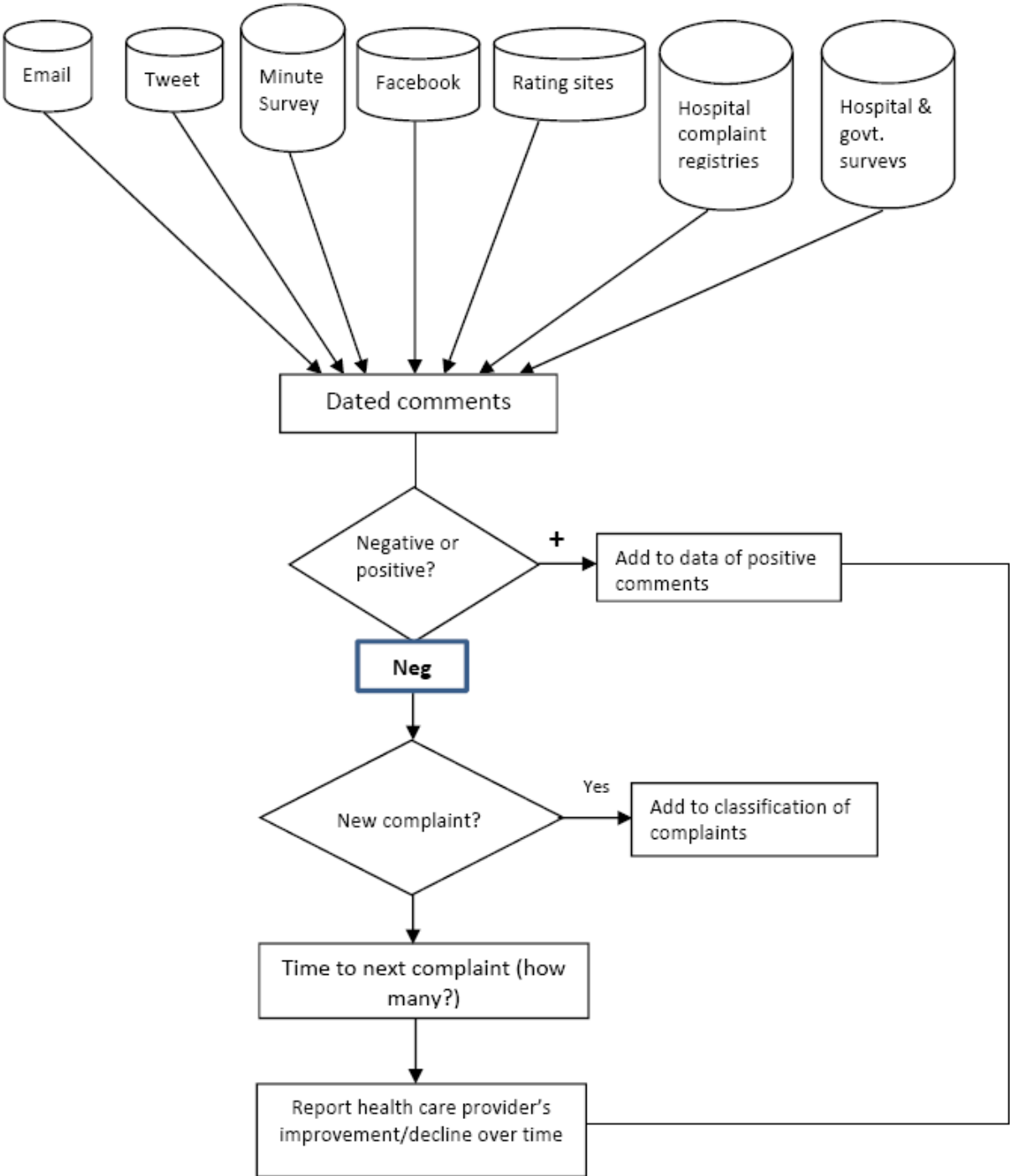
- [Example list](#) //LR 5: Intensifier rule
- Less obvious examples:
 - *Accumulation* can also act as an intensifier, as in ‘an accumulation of disasters’ = many disasters, which would be scored as *many* +2 and *disasters* -2 = -4.
 - *Thousands, thousandth, hundreds, hundredth*, in context, they are intensifiers used figuratively and not literally as a number.
 - *Billion* and *billions* used as a hyperbolic intensifier, +2.
 - *Astronomical* and *astronomically* +2 (as in *hugely*)
 - *Countless*, morphologically negative, but intensifier.

I was left in the waiting room in front of countless other people who were seen from pillar to post. After many calls and countless times of being told that i have a conclusion to my health problem as countless trips to my GP have so far although even at the end I was finding countless seeds. Makes me wonder be able to get the results, have made countless calls over the last 7 months this expert has proven them wrong on countless occasions. A elderly, open, having her cuddly toy near her. Countless things, and this knowledge is than a dirty waiting room where countless people have sat, sneezed, detail. The monitoring has involved countless blood tests, scans and 2011. That's 9 months! After 9 months, countless visits to A&E, GP, referrals, choice whether to proceed. I feel sure countless others have had bad called specialist I spoke to. I have heard countless horror stories about crisis 5 months now. I have been to my GP countless amount of times and I have yet happened. My mother has enquired countless times about my health and times a day and went back to theatre countless times to have it cleared out. after my baby's birth. Having read the countless negative media stories about did not and would not deal with. I went countless times about constant flare heart bypass surgery procedures, and countless visits to her wonderful , with my whole body burning and with countless attempts to control this by see what was happening. We received countless apologies for lack of care,

4. CONCLUSION AND FUTURE RESEARCH

- Words enter into relationships with other words
- Proposed solution is that we need to be looking at different levels of text: keywords (single words), n-grams, sentences, text, and a corpus of texts
- Linguistic & semantic problems
- Potentially enormous benefits

A system for monitoring health service quality



REFERENCES

- Alemi, F., Torii, M., Clementz, L., & Aron, D. C. (2012) 'Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments', *Quality Management in Healthcare*, 21(1), 9-19.
- Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography', *Computational Linguistics*, 16(1), 22-29.
- Feldman, R. (2013) 'Techniques and applications for sentiment analysis', *Communications of the ACM*, vol. 56, no. 4, pp. 82 [Online]. DOI: 10.1145/2436256.2436274
- Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., and Donaldson, L. (2013) 'Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online', *Journal of Medical Internet Research*, 15(11).
- Liu, B. (2010) 'Sentiment analysis and subjectivity', in Nitin Indurkha and Fred J. Damerau (ed) *Handbook of Natural Language Processing*, Second edn. Boca Raton, FL, CRC Press, Taylor and Francis Group.
- Pang, B. and Lee, L. (2008) 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1-135.
- Taboada, M., J. Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) 'Lexicon-based methods for sentiment analysis', *Computational Linguistics*, vol. 37, no. 2, pp. 267-307.
- Xia, L., Gentile, A.L., Munro, J. and Iria, J. (2009) 'Improving patient opinion mining through multi-step classification', [Online]. Springer.