EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 381–395
CILC2016. 8th International
Conference on Corpus Linguistics

A Corpus-Driven Approach to Sentiment Analysis of Patient Narratives

Keith Stuart, Ana Botella, and Imma Ferri
Universitat Politècnica de València, Valencia, España
kstuart@idm.upv.es, apbotell@idm.upv.es, imfermi@idm.upv.es

Abstract

This paper describes the linguistic analysis of a corpus of patient narratives that was used to develop and test software to carry out sentiment analysis on the aforementioned corpus. There is a growing body of research on the relationship between sentiment analysis, social media (for example, Twitter) and health care, but less research on sentiment analysis of patient narratives (being longer and more complex texts). The motivation for this research is that patient narratives of experiences of the National Health Service (NHS) in the UK provide rich data of the treatment received.

The corpus threw up some unexpected results that may be of benefit for researchers of sentiment analysis. The linguistic problems encountered have been divided into three sections: the noisy nature of large corpora; the idiomatic nature of language; the nature of language in the clinical domain. This article gives an overview of the project and describes the linguistic problems that arose out of the project, which tried to find a means to automate the analysis of patient feedback on health services.

1 Introduction

Sentiment analysis is the computational study of evaluative expressions in text. There are many advantages (security reasons or product/service evaluation) to be derived from information systems that are able to extract information about people's appraisal, opinion, and sentiment about an event, individual, organization, product, service or topic. Sentiment analysis is the task of finding people's opinions about specific entities in text, a technique to classify people's opinions. Research in sentiment analysis has focused mainly on two problems: detecting whether the text is subjective or objective, and determining whether the subjective text is positive or negative. An important concept in sentiment analysis is semantic orientation, which refers to the sentiment polarity (positive or negative) and sentiment strength of words, phrases, or texts. It is often the goal of sentiment analysis to find the semantic orientation of texts (Taboada *et al*., 2011; Mohammad *et al*., 2013).

The objective of this paper is to describe and give examples of the complexities of language when trying to carry out Sentiment Analysis. The paper is based on a research project in Corpus Linguistics

and Sentiment Analysis (SA) applied to the National Health Service (NHS) in the UK. This project[*] tried to automate the analysis of a corpus of patient narratives of their experiences with the NHS (UK). The end product of the project was software for Sentiment Analysis of a corpus of Patient Narratives. However, we shall be concentrating on the linguistic and semantic aspects of the project. In particular, we shall be analyzing three aspects: the noisy nature of large corpora, the idiomatic nature of language, the nature of language in a specific discourse or knowledge domain, in our case, the clinical domain.

2 Sentiment Analysis

In general, opinions can be expressed on anything. Therefore, the applications of SA are many. As (Feldman, 2013: 82) affirms,

> There is a huge explosion today of 'sentiments' available from social media including Twitter, Facebook, message boards, blogs, and user forums. These snippets of text are a gold mine for companies and individuals that want to monitor their reputation and get timely feedback about their products and actions. Sentiment analysis offers these organizations the ability to monitor the different social media sites in real time and act accordingly. Marketing managers, PR firms, campaign managers, politicians, and even equity investors and online shoppers are the direct beneficiaries of sentiment analysis technology.

From a less commercial point of view, one could add to this list of 'beneficiaries of sentiment analysis technology' the health and educational services, the police, the judiciary and any counter-terrorist organization.

The main reason for doing this kind of research is that Sentiment Analysis applications have the potential to develop into important social technology for large scale collaborative decision-making and policy-making. These applications can act as monitoring systems of possible social discontent, by detecting early feedback from citizens. One such example is feedback from citizens who have encountered a problem in the health services.

However, 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.

2.1 What are the challenges of Sentiment Analysis?

The main challenge in SA is the complexity of language. This can be illustrated through an examination of negation. Negation has a major impact on the contextual polarity of opinion words and texts. Negation words† or phrases, such as *never, not, no, none, nothing, neither, nor* can reverse the polarities of the opinion words. Similarly, language patterns such as "stop + vb-ing", "quit + vb-ing" and "cease + to-inf vb" can express negation and a negative evaluation but it depends on the social context of the text.

a) The latest iPhone is great (positive) → The latest iPhone is ***not*** great. (negative)

b) My iPhone stopped working (negative) → The medicines worked. The tumour stopped growing. (positive)

[*] This project was carried out in collaboration with Patient Opinion: https://www.patientopinion.org.uk/info/about. More information about the project can be found here: https://patientopinioncorpus.wordpress.com.

[†] These also include *haven't, hasn't, hadn't, can't, couldn't, shouldn't, won't, wouldn't, don't, doesn't, didn't, isn't, aren't* etc.

More specifically, in our corpus, there were complex linguistic problems with sentences such as the following:

1. Admission was ***haphazard*** although the staff were ***very nice*** but ***very busy***. (negative–> positive–>implicitly negative)
2. ***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)

In sentence 1, the adjectives *haphazard, nice, busy* bear the weight of the evaluation. However, both *but* and *although* are sentiment shifters that alert the reader to a change in the semantic orientation of the text. Furthermore, *very nice* seems to be in contrast to *very busy*. The word *busy* is not intrinsically negative but in the context of the NHS (and our corpus) it usually is. *Busy* is equated to being understaffed, not able to devote sufficient time to the patient and a general sensation that people are rushing around in the medical profession.

Sentence 2 has greater complexity as there is a lot of implicitly negative evaluation going on. 'I would have liked more information' means I didn't get the information that I wanted which is emphasized further by 'information … is rather vague'. Modal perfects like *would have liked* have been given special treatment in the linguistic analysis of our Patient Opinion Corpus (POC). There are all kinds of social interpretations that this short text has for an alert reader ('I need assurance', 'I live alone').

In both sentences, there are intensifiers (*very, rather*). In our corpus, we found 169 different kinds of intensifiers (whether they be amplifiers or downtoners). We will return to the question of intensifiers below as there are intensifiers that are not so easy to detect automatically.

As we said above the main challenge in SA is the complexity of language. This complexity arises from 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***. (sentiment?)

Patient Opinion Corpus:

- This is my first experience for a ***long time*** with the NHS, and 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)

And, if the reader had more context, what would their assessment be of the sentiment expressed?

- …but I was aware that this was due to the hospital being extremely busy so there's not anything else that could be done.

2.2 Sentiment analysis in the clinical domain

There is a growing body of research on the relationship between sentiment analysis, social media and health care (Verhoef *et al.,* 2014). However, there are several weaknesses in these studies:

1. Most studies focus on health care rating sites (a rather simplistic form of sentiment analysis of patient experience).
2. They report on extremely limited forms of communication. For example, (Timian *et al.,*

2013) investigated the number of "likes" on the Facebook pages of 40 American hospitals. In other words, they have a thumbs up or thumbs down style of sentiment analysis research.

3. Some report on the use of Twitter as a source of information about quality of care (Greaves *et al*., 2014), although these short, unstructured messages contain minimal information.

There is much less research on sentiment analysis of patient narratives, which are much longer and more informative texts. Xia *et al*. (2009: 79) hypothesize that polarity is not statistically independent from topic and that learning topic-specific polarity classification models can help improve the accuracy of polarity classification systems. More formally, they propose that a document collection is $D = D_T \cup D_P$, a set of topics T, and training labels given by $L_T: D \rightarrow T$ and $L_P: D \rightarrow \{-1, 1\}$, for topic and polarity respectively. Using a standard multinomial Naïve Bayes approach for learning the classifiers, they propose the following steps:

1. learn a topic classifier from D_T using labels L_T, by approximating a classification function of the form $f_T: D \rightarrow T$;
2. Apply f_T to D_P, thereby splitting D_P by topic, that is, creating sub-datasets

   $D_{Pt} = \{ d \in D_P: f_T(d) = t\}, \forall t \in T$;
3. For each $t \in T$, learn a polarity classifier from D_{Pt} using labels L_P, by approximating a classification function of the form $f_{Pt}: D \rightarrow \{-1, 1\}$.

In the Naïve Bayes probabilistic framework, a document is modelled as an ordered sequence of word events drawn from a vocabulary *V*, and the assumption is that the probability of each word event is independent of the word's context and position in the document. This is something we would not agree with and have introduced into our analysis the possibility of taking word context into account. Another criticism of this early research is that they only analysed 1200 patient comments, a relatively small corpus compared to our 50,000 patient comments.

Greaves *et al*. (2013) applied machine learning techniques to 6412 online comments from the large number of free-text comments on the UK NHS Choices website. They used Weka data-mining software, which allowed them to examine the data through sentiment analysis. These comments are matched with the users' own quantitative ratings of the service, which meant the authors had the opportunity to measure the accuracy of natural language processing methods against the patient's own assessment. Similarly, they used the results of the NHS programme of patient experience measurement via a national survey of hospital inpatients. Using these data sources, they could compare sentiment analysis of patient comments on the NHS Choices website with traditional patient surveys carried out by the NHS itself. More particularly, they tried to predict whether a patient would recommend a hospital, whether the hospital was clean, and whether they were treated with dignity from their free-text descriptions.

According to their data, they were able to predict patients' rating of their care from their free-text comments with an accuracy of 81% for hospital cleanliness, 84% for treatment with dignity, and 89% for overall recommendation. They observed mild to moderate associations between their machine learning predictions and patient survey quantitative responses for the three categories examined (Spearman rho 0.37-0.51, $P<0.001$ for all).

The prediction accuracy that they achieved using machine learning processes suggests that it is possible to predict from free text, a reasonably accurate assessment of patients' opinion about different performance aspects of a hospital. They also suggest that it might be possible to monitor the so-called online cloud of patient experience in real-time and by doing so harness the value of patient opinion.

Alemi *et al.* (2011) were also interested in harnessing patient opinion and utilized a publicly available sample of 995 comments from the Rate My MD site (http://www.ratemds.com/). Interestingly, these comments cover a 5-year period. A corpus of patient experience will be richer data if the size of the corpus is by default large and is the result of data that covers a period of time. Precisely, this research paper highlights the idea of using the average number of visits-to-next complaints as an overall measure of satisfaction. In other words, the longer it takes to receive a complaint about a health care provider, the lower the rate of dissatisfaction. This can only be done if one has a corpus covering a period of time as in our study (our corpus of texts range from 2008 to 2014). Theoretically, we could say on X date a complaint is registered about X unit in X hospital and this unit can be tracked over time. However, for this, one would need a large quantity of reliable data. We return to this question in the conclusion section of the paper.

Most researchers use machine learning techniques (whether supervised or unsupervised) to carry out SA on texts. In figure 1 below, neural networks are given as an example of both a supervised and unsupervised learning method. However, machine learning solutions include a variety of techniques such as Support Vector Machines, Naïve Bayes, Maximum Entropy for supervised learning involving labelled trained data. Unsupervised learning techniques without trained labelled data include clustering. Among neural network models, the self-organizing map (SOM) and adaptive resonance theory (ART) are commonly used unsupervised learning algorithms.

Our research focusses on a corpus-driven approach to SA using both semantics, language patterns and statistics. Figure 1 summarizes different approaches to SA.

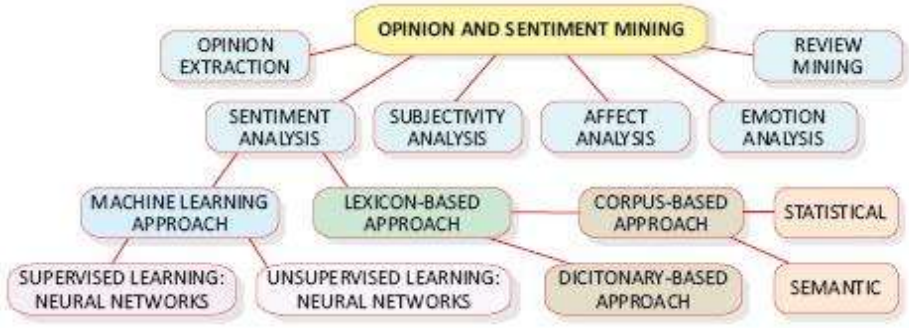

Figure 1: Problem space and techniques in Sentiment Analysis

3 Patient Opinion Corpus: collecting and analyzing the data

The basic data set was a corpus of patient narratives (www.patientopinion.org.uk) giving feedback about treatment for a medical condition within the NHS. Patient Opinion is the UK's leading independent non-profit feedback platform for health services. Patient Opinion kindly provided us with an API key. We downloaded 50,000 XML documents and converted them into txt format. Plain text

format is best for pre-processing and analysis of linguistic data (for more information on the project: https://patientopinioncorpus.wordpress.com/).

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

Table 1: POC (Patient Opinion Corpus)

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 (the words that have the greatest impact on sentiment scores) were scored on a scale of intensity in the following manner. There is a fair amount of subjectivity in the process.

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

Table 2: Evaluative words in POC

What we have discussed so far is a bag of words approach. But, a bag of words approach of a document d with all words scored for a sentiment score s can only get us so far. Language is far more complex and dynamic.

Words enter into relationships with other words. These in turn form phrases/clauses that form sentences that form texts. The 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. The solution our software tries to implement is multi-level Sentiment Analysis (figure 2).

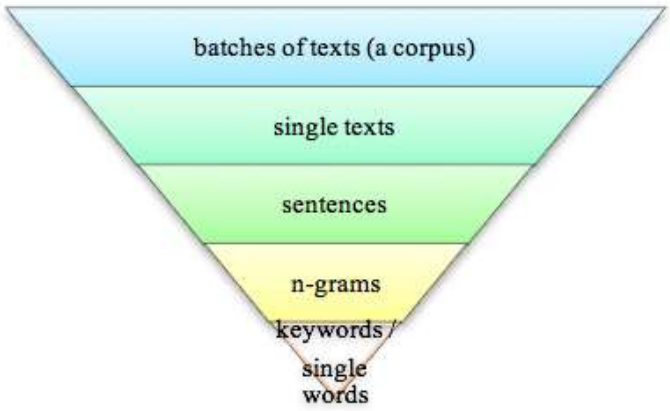

Figure 2: Multi-level Sentiment Analysis

This can be visualized through a screenshot of the interface of the software Sentiscore (figure 3).

Figure 2: Interface of the software Sentiscore

Just as we give weights to a word, we give weights to sentences. This is done by multiplying word scores. We then divide the sentence weight by the total number of words to give us the sentence mean which decides whether a sentence is positive or negative. In many Sentiment analysis systems, a text is positive or negative depending on how many positive or negative sentences it contains. In other words, the software counts the number of positive and the number of negative sentences and,

depending on which is greater, gives a positive or negative orientation to the text in question. I do not use this system rather I use a total added score for text orientation. However, as we have already stated, a bag of words approach only takes us so far, words are dynamic variables and language is complex. Some of these linguistic complexities are illustrated and introduced into the system as linguistic rules. In figure 3, we have introduced a tab for linguistic rules and the software has the option to apply different rules. Some of the rules are the following: LR1 Negative Idiom Rules, LR2 Modal Perfect rule, LR 3 Reduced distress rule, LR 4 Increased distress rule etc. The software also incorporates some more basic general rules such as reversal of polarity because of negation which are not optional. It is precisely through the study of these rules that we encountered some linguistic problems that have not been previously analyzed and discussed in the literature on SA and that we have divided into three sections: the noisy nature of large corpora, the idiomatic nature of language, and the nature of language in the clinical domain.

4 Linguistic and Semantic Problems

4.1 Noisy nature of large corpora

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, acctualy, actualy, acctually*, etc.), informal language *(C'mon, geez*, etc. used for emphasis; *blimey, cor, crikes, crikey, effin*), abbreviations (WTF, WTH) 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	ADELAY	AWEEK	ACHEST	AFOREST
ADAY = a day	AFRACTURE	ABAD	ACOMPLETE	AGOWN
AGREAT = a great	AFURTHER	ABADLY	ACUBICLE	AHORRIBLE
AGOOD = a good	APERSON	ABEAUTIFUL	ACUP	AJOKE
ANURSE = a nurse	ASPECIAL	ACAST	ADIFFICULT	ALACK
ABUSY = a busy	ATICKET	ACAT	ADR	ALARGE

Table 3: Example of noise in the data (indefinite article plus adjective/noun typed together)

The whole database is bombed with typos, and phonetic misspellings (i.e. the writer comes up with a spelling for an unknown word that s/he knows and uses in oral language but can't write), so something like *cry's* could well be *cries*, that *curtecy* and *curtios* stands for *courtesy* and *courteous*, etc. However, one has to almost admire the creativity of the patients that came up with *appsalutly* and *abserlootly*. There are problems with *hell* and *well*. These are sometimes misspelled versions of *he'll* and *we'll*.

Noise may turn out to be very troublesome when the corpus is not cleaned. Errors in the corpus may induce biases and distort the reliability of analytical results. Patient narratives are a noisy channel for obtaining data about the health services but the data is rich. There may even be the need for having procedures in place to recognise an 'unreliable' narrator in the data.

4.2 Idiomatic nature of language

The idiomaticity of text is a problem for automating SA because it implies that text may have a large number of preconstructed multi-word combinations (multi-word expressions) which behave like a single semantic unit and may be positive, negative or neutral as far as their evaluative propositional

content is concerned. They are problematic because they have meanings that are not predictable from the properties of the individual lexemes in the multi-word expressions. In other words, you cannot predict their meaning by aggregating their different parts. Sinclair (1991: 110) expresses the same idea by proposing the idiom principle that a language user has available to her/him a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. He contrasts the idiom principle with the open choice principle, which is the assumption that practically each position in a clause offers a choice.

More specifically, questions of idiomaticity in text beg some of the following questions:

1. What is the average proportion of these multi-word expressions in texts?
2. How do multi-word expressions interact with words combined according to the open choice principle?
3. What is the distribution of multi-word expressions and strings of non-preconstructed multi-word combinations throughout text?

Presumably, it is possible to relate certain types of multi-word expressions to certain types of text. The frequency of multi-word expressions might depend on whether the text is spoken or written. In the POC corpus, although the genre is a narrative written text, the narratives are often written, as if spoken, which might explain the large number of these multi-word expressions. In particular, we found many negative idiomatic phrases.

burst in to tears	-5,00	horror stories	-4,00
cannot stand	-4,00	in a mood	-3,00
concerned with	-2,00	in a state	-4,00
do nothing for	-2,00	in tears	-4,00
doom and gloom	-4,00	knocked out	-4,00
fed up	-3,00	on edge	-2,00
find fault	-3,00	reduced to tears	-5,00
freaked out	-4,00	shake up	-4,00
frightened to death	-5,00	stressed out	-3,00
grim looking	-4,00	take out on	-2,00
had enough of	-3,00	taken aback	-3,00
hit the roof	-4,00	trouble maker	-3,00

Table 4: Some examples of negative idioms with their sentiment score in the POC corpus

The oral nature of the texts in POC has also led to many words being slang or swear words. Some examples are given below:

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

Another aspect of the idiomatic nature of language is the frequency of the use of intensifiers in the POC corpus. In our corpus, we found 169 different kinds of intensifiers but some intensifiers are not easy to detect and have an idiomatic nature. *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

equals -4. *Thousands*, *thousandth*, *hundreds*, *hundredth*, in context, are intensifiers used figuratively and not literally as a number. *Billion* and *billions* are used as a hyperbolic intensifier, +2. *Astronomical* and *astronomically* +2 are often used as in the sense of meaning *hugely*. *Countless* is morphologically negative, but acts as an intensifier and more often than not with negative connotations.

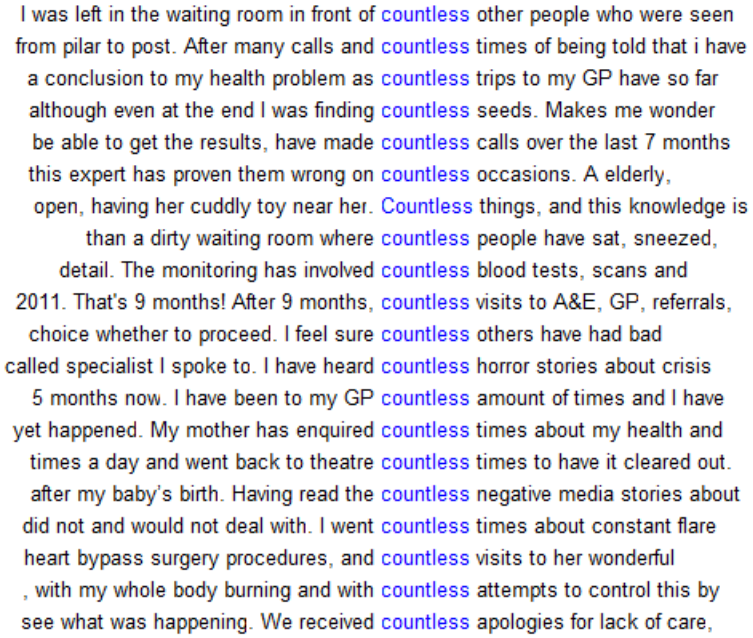

Figure 3: Concordance lines of *countless* as an intensifier

In the POC corpus, we have also found that idiomatic use of polysemous words is another problematic area for automating SA. Word sense disambiguation has been recognized as being difficult for Natural Language Processing (NLP) applications in general and getting one's system to work reasonably well on any given multi-sense word will be very hard indeed. One can try several methods to determine the right sense of a word. In principle, Part-of-Speech (POS) tagging will reduce the number of candidate senses. Probability of word co-occurrence might be another method. Cosine similarity to other words is another technique. However, the disambiguation of polysemous words is very difficult as can be shown from a few examples from the POC corpus.

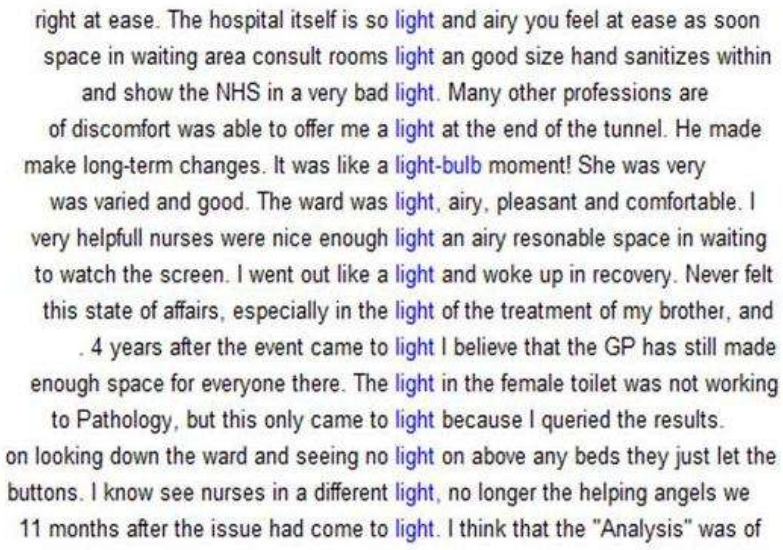

Figure 4: Concordance lines of the polysemic word *light* in the POC corpus

As can be seen in the concordance lines for the polysemic word *light*, the situation is further compounded as *light* can express positive meanings as in *light and airy* or *light at the end of the tunnel*, neutral meanings such as *the light in the female toilet*, and negative meanings such as *show the NHS in a very bad light, the issue had come to light*. One can also note how polysemous words tend to enter into multiword expressions. Two more examples are given below with the words *ball* and *bare*. The second example is particularly illustrative of the difficulties of accurate SA. *Bare* is not only used in a polysemic manner but is also misspelt. The writer meant to say *bear* as in the meaning of withstanding something unpleasant and/or painful. So we have here the polysemic problem mixed up with a noise problem because the writer does not know how to spell the word correctly.

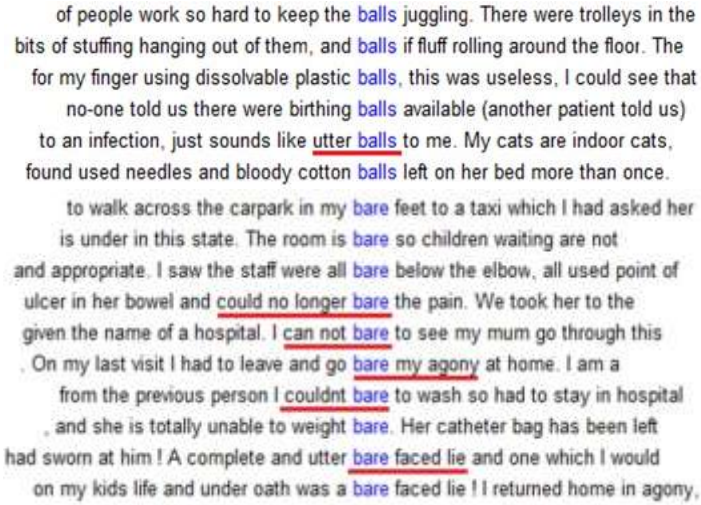

Figure 5: Concordance lines of the polysemic words *balls* and *bare* in the POC corpus

There are many more issues related to the idiomatic nature of language: animal imagery: *cow, leech, tadpole*; animal onomatopoeia: *bark, bleat, bray, cackle, caw, coo, howl, yap, yelp* (they are all in the corpus); figurative language; irony; sarcasm. Figurative language, in particular, is very difficult. When patients use language creatively, it is harder to make their sentiment fit into a score or even to know how to give a sentiment score to expressions like a *cinderella service* or *circus act*. Below are some examples from the POC corpus.

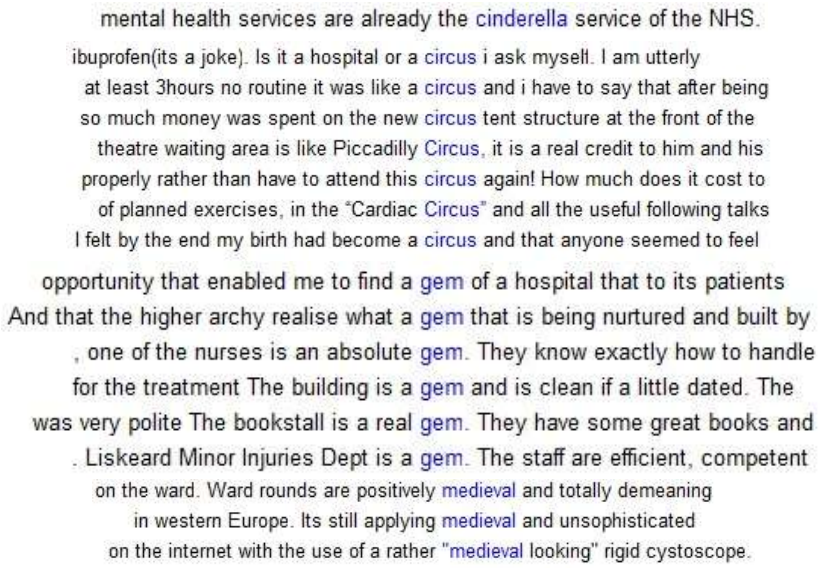

Figure 6: Figurative use of language in the POC corpus

The large quantity of idiomatic expressions and concordance lines from corpora such as our POC corpus confirm their numerousness in text. This could be considered a major problem for SA or a window of opportunity for linguistic analysis from the point of view of pattern recognition.

4.3 Nature of language in the clinical domain

Regardless of whether the patient has had a positive or a negative experience with the service received at the hospital, one can identify some context-specific semantic sets which tend to be loaded with negative sentiment:

- Pathologies (*cancer* 2103, *stroke* 1040, *cardiac* 525, *tumour* 317 etc.)
- Symptoms (*pain* 10970, *infection* 2188, *broken* 1250 etc.)
- Treatments (*surgery* 8759, *operation* 7506, *injection* 536 etc.)

Then there are other sets of words which are loaded with positive sentiment:

- Care, Caring, Comfort, Compassion, Consideration, Courtesy, Help, Kindness, Reassuring, Recovery, Sensitivity, Support
- Reassuring (1390), Reassured (794), Reassurance (388), Reassure (263), Reassuringly (26)

The problem here is that, if we score these words negatively, we are not fulfilling the objectives of the research. The objective of the research is through the analysis of patient narratives to discover patient sentiment towards the National Health Service. We want to extract information about their opinion of the service being provided. If the service *stops the pain*, this is indeed positive, but *cancer/pain* are not in themselves a negative attribute of the service. They are the reason why doctors, hospitals and nurses exist but the health service in principle does not cause the cancer.

This observation points to a particular aspect of the semantics of the clinical discourse domain. Patients are in hospital to be cured and, hopefully, their health is improved. However, patients are often distressed by their situation and this is reflected in the corpus. We introduced into our scoring system two semantic rules:

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

… to sit there for over 2 hours waiting to be called, **my pain gradually increasing** all the time.
I went back to my husband's bedside feeling **increasingly uncomfortable with the situation**.
… my labour was induced as **I had become increasingly worried** about reduced movements…
When I finally got the results it turned out that I had a **tumour growing** on my spinal cord.
Everyday my mum tells me she is suffering and in pain and as the **tumour grows** the pain gets worse.
The A&E doctors, nurses and specialist doctors were all fantastic and arranged for emergency surgery to **stop the bleeding** and a blood transfusion.
Mr Pembridge carried out the procedure and experienced difficulty in **stopping me losing blood**, but again fantastic work by him to **stop the bleeding** and finish off the procedure.
I was given a spinal block and from then on **the pain stopped**…

In the corpus, there are 1828 examples of stop X (resulting in some improvement) which include 91 examples of the pattern 'stop the pain' and 66 examples of the pattern ´stop the bleeding´.

The semantics of the clinical discourse domain is such that the polarity of an outcome (of a health problem) is often determined by how change happens. If a bad thing (the probability of mortality, serious illness or pain) was reduced, then it is a positive outcome. If the bad thing was increased, then the outcome is negative. Therefore, a crucial issue for SA of texts from the clinical domain is identifying the polarity of clinical outcomes. However, this should not be confused with the perception of how a patient was treated.

5 Conclusions and Future Research

Acute, critical urgent, bloody are all words (and there are many more) that take on particular meanings within the clinical discourse domain in the POC corpus. They are text instantiations of the socio-sanitary context of the NHS. There are many different ways social context enters and shapes the texts that make up a corpus. This often means your corpus can be messy and dirty, as we have already seen above. However, regardless of how messy or dirty your corpus is you need to understand all its nuances. Knowing the linguistic data in the corpus is a first step to a successful analysis effort. This means that the first step in an analysis effort is to establish consistent processes to clean the linguistic data.

What also emerged from the research that there is no single tool that allows you to perform all of your corpus data analysis tasks. Many different tools exist, and each tool has a specific purpose. In order to be successful analyzing corpus data, one should have access to the tools one needs and also be able to configure these tools as needed. Providing one fixed tool or trying to design and develop one tool to perform all task is unrealistic and unreasonable. In our attempts to design and develop a tool, we combined a lexicon approach (a database) with linguistic rules. When we came to analyze the POC corpus, we also used a POS tagger and a Semantic tagger. It has become clear to us that one needs a set of tools to successfully carry out Sentiment Analysis on a corpus of texts as language

functions on many levels simultaneously. A central aspect of automated language analysis is the requirement of an algorithm to carry out a search of a complex and multidimensional problem space.

A very simplistic way of defining this problem space is to say that words enter into dynamic relationships with other words. These, in turn, dynamically and interactively form phrases/clauses that form sentences that form texts. The 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. Keywords such as *staff, doctors, nurse, admissions, appointment, treatment, care* are important because we want to know what the client (a patient) feels about these key actors and processes.

For fine-grained sentiment analysis on a sentence and sub-sentence level, we need to analyze words, word phrases (n-grams) and sentences. To achieve better modelling of compositional sentiment (the collective sentiment of patients), we need to look at all levels including whole texts and batches of text.

There are linguistic and semantic problems that should not be underestimated. In this article, we have presented a few of these problems and have categorized into three groups: the noisy nature of large corpora; the idiomatic nature of language; the nature of language in the clinical domain. Despite the linguistic difficulties involved, texts are very rich data and there are potentially enormous benefits from this kind of research.

You may want to look at sentiment development over a period of time (how people's feelings change about something over time). In the case of our POC corpus, you may want to know how many complaints have been made about a particular unit at a particular hospital over a period of five years. As the data was collected over a six-year period, this kind of analysis is possible with our Patient Opinion Corpus. Information gathered through sentiment analysis on a corpus of patient narratives could ultimately make a difference to the running of hospitals and the wellbeing of patients.

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.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82-89.

Greaves, F., Laverty, A., Ramirez-Cano, D., Moilanen, K., Pulman, S., Darzi, A., Millett, C. (2014). Tweets about hospital quality: a mixed methods study. *BMJ Quality & Safety, 23*, 838–846.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of Medical Internet Research, 15*(11), e239.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436-465.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Taboada, M., J. Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267-307.

Timian, A., Rupcic, S., Kachnowski, S., & Luisi, P. (2013). Do patients "like" good care? Measuring hospital quality via Facebook. *American Journal of Medical Quality, 28*, 374-382.

Verhoef, L. M., Van de Belt, T. H., Engelen, L. J., Schoonhoven, L., & Kool, R. B. (2014). Social media and rating sites as tools to understanding quality of care: a scoping review. *Journal of Medical Internet Research, 16*(2), e56.

Xia, L., Gentile, A.L., Munro, J. & Iria, J. (2009). Improving patient opinion mining through multi-step classification. *TSD '09 Proceedings of the 12th International Conference on Text, Speech and Dialogue* (pp. 70-76). Pilsen, Czech Republic: Springer-Verlag.