Towards Objective and Reproducible Study of Patient-Doctor Interaction: Automatic Text Analysis Based VR-CoDES Annotation of Consultation Transcripts

Charlotte Birkett† and Ognjen Arandjelović‡ and Gerald Humphris‡
†School of Computer Science ‡School of Medicine
University of St Andrews
Scotland, United Kingdom

Abstract—While increasingly appreciated for its importance, the interaction between health care professionals (HCP) and patients is notoriously difficult to study, with both methodological and practical challenges. The former has been addressed by the so-called Verona coding definitions of emotional sequences (VR-CoDES) — a system for identifying and coding patient emotions and the corresponding HCP responses – shown to be reliable and informative in a number of independent studies in different health care delivery contexts. In the present work we focus on the practical challenge of the scalability of this coding system, namely on making it easily usable more widely and on applying it on larger patient cohorts. In particular, VR-CoDES is inherently complex and training is required to ensure consistent annotation of audio recordings or textual transcripts of consultations. Following up on our previous pilot investigation, in the present paper we describe the first automatic, computer based algorithm capable of providing coarse level coding of textual transcripts. We investigate different representations of patient utterances and classification methodologies, and label each utterance as either containing an explicit expression of emotional distress (a ‘concern’), an implicit one (a ‘cue’), or neither. Using a data corpus comprising 200 consultations between radiotherapists and adult female breast cancer patients we demonstrate excellent labelling performance.

I. INTRODUCTION

Communication, and interaction in general, with a patient is increasingly appreciated as an important aspect of health care provision [17], [27]. Most obviously, it is crucial that information is presented in a manner which is both clear and readily understandable by laymen, and comprehensive in the sense that it empowers individuals to make decisions appropriate to their needs. A subtler but no less important aspect of communication which has come into prominence more recently concerns the tailoring of interaction to meet a patient’s emotional needs [20], [27]. There is a growing corpus of evidence that demonstrates remarkable differences not only in the short, but also in the medium to long term, on patients’ mental condition [23], [22], primary outcomes [10], [14], and numerous others [1], [11], [15], [26].

However, studying interaction between health care providers (HCPs) and patients is a challenging task. The first difficulty is a methodological one. In particular, as always in the course of scientific inquiry, the objectivity and repeatability of the process are of paramount importance. Yet, it is far from obvious how this can be achieved while at the same time ensuring that the phenomenon of interest is not excessively abstracted so as to lose on clinical significance and insight it can offer. As we explain in some detail in the next section, this aspect of the challenge has been addressed in the past with the introduction of the so-called Verona coding definitions of emotional sequences (VR-CoDES). In brief, VR-CoDES is a system for the identification and categorization of types of emotional expressions by patients, as well as the corresponding HCPs’ responses. Notwithstanding the sizable body of evidence demonstrating the repeatability of VR-CoDES categorization [5], [12] and the novel insights this framework has already produced, a major practical limitation is posed by the complexity of the coding hierarchy, the need to train researchers in its use, and the labour involved in labelling consultation recordings.

The present work is aimed at addressing a practical limitation of VR-CoDES use. Our overarching goal is to develop computer based tools which can assist in the labelling of patient-HCP interactions. Specifically herein we wish to assess the feasibility of using textual transcripts of consultations only (as opposed to audio recordings, or both audio recordings and textual transcripts) in the labelling process. In particular, we address the problem of labelling a patient’s utterance (verbal communication, uninterrupted by the HCP, which comprises a single or multiple sentences) as containing (i) an explicit concern, (ii) an emotional cue, or (iii) neither. Thus, in addition to addressing the specific issues of patient-HCP communication, the present work contributes to the overall effort in the use of modern computer science in assisting health care provision [4], [7], [24], [25].

II. FORMALIZING EMOTION COMMUNICATION ANALYSIS

The Verona coding definitions of emotional sequences is a framework for extracting and semantically categorizing salient aspects of patient-HCP provider exchanges in the context of emotion analysis. The framework covers both emotion expression by patients and the corresponding responses by HCPs. A major challenge faced by HCPs in addressing patients’ emotional needs is posed by the observation that the expression of emotion is usually subtle and non-explicit [27]. For instance, previously stable patients may complain about new physical symptoms that might suggest underlying emotional distress [21]. Hence, at the first, coarsest level of hierarchy, patient expressions of emotion are categorized...
as being either explicit or implicit. In the context of VR-CoDES the former are termed concerns, and the latter cues (observed much more frequently than concerns, as we discuss in Section IV). At increasing depths of the coding hierarchy, these labels are refined further. For example cues are categorized into seven groups: [cue a] the use of vague or unspecified words to describe emotions (e.g. “that for me was oo, a big deal”), [cue b] verbal hints to hidden concerns (e.g. “but I’m just aware that those mood swings have come back, and I think, oh dear”, [cue c] words or phrases which emphasise physiological or cognitive correlates of unpleasant emotional states (e.g. “…well, I’m waking up during the night”), [cue d] neutral expressions that mention issues of potential emotional importance (e.g. “I feel as if I’ve never really had a break from the hospital”), [cue e] repetition of a previous, neutral expression (e.g. “It’s just nice to sort of know what to look out for”), [cue f] non-verbal expression of emotion (e.g. “The past couple of days have just...-sigh”), and [cue g] clear and unambiguous expression of an unpleasant past emotion (e.g. “I was really, really frightened”. Comprehensive detail with examples can be found in the relevant previous work [13], with manuals available on the European Association for Communication in Healthcare (EACH) website (http://www.each.eu/).

III. METHODS

In this section we describe our data collection protocol and summarize the machine learning techniques evaluated using the collected data.

A. Participants

Our data was obtained using a cohort of adult female breast cancer patients aged 28 to 85 years (average 58 years, standard deviation 11.3 years) who were undergoing radiotherapy at the Edinburgh Cancer Centre, Scotland. Audio recordings of one-on-one consultations from an observational cohort study of patients and radiotherapists were collected using studio quality equipment. In total, 200 consultations involving 91 patients and two therapeutic radiographers were collected. Participants were consenting volunteers, had no known psychiatric conditions and were able to communicate in English. Ethical approval was obtained from the University of St Andrews ethics committee. Audio recordings were subsequently professionally transcribed to produce semi-structured textual transcripts used as input to our algorithms.

B. Experimental data

Behavioural coding was completed by two researchers. Cue F (non-verbal emotion) and the ‘silence’ responses were not coded due to an inability to identify them in audio recordings without visual cues. Demographic and treatment data were also collected, including patient age, radiotherapy type (three types), chemotherapy use (yes or no), degree of cancer recurrence fears before study (on a 16 point scale), self-reported rating of general ‘health state’ (1–100 scale), living situation (alone or not), consultation number (first, second, third, or fourth), and the consultation duration. Patient utterances, in accordance with the VR-CoDES standard, were labelled as either containing a specific concern or a cue (of which there are 6 types – see the original body of publications for additional background information).

Evaluation was performed using the standard 5-fold cross-validation protocol, in an effort to minimize the potential of overfitting. Performance was assessed using several widely adopted measures: the average classification accuracy, the area under curve (AUC) of the recall-precision characteristic, the F-score, and the $\kappa$ statistic.

C. Utterance representations

Automatic, computer based analysis of textual documents is now a well-established field within machine learning [9]. However, it is important to appreciate the specific nature of our data which makes the problem at hand rather different from those in the existing literature. In particular, utterances in patient-HCP consultations tend to be rather short, usually comprising a single or a couple of sentences. This aspect of our data makes it very different from the corpora of longer documents for which the state of the art tools have been developed [9]. All else being the same, shorter pieces of text contain less information and have already been recognized as not being well suited to the types of approaches used generally [6]. On the other hand, unlike corpora of short texts used previously (such as Twitter messages), our corpus is much more constrained by its very nature. For example, it can be expected that the vocabulary used in patient-HCP consultations is much smaller than that used in semantically virtually unrestricted exchanges. In other words, while it is true that chunks of text that we are dealing with here are short in the absolute sense, this is probably misleading in that relative discriminative information content is denser.

1) Term abstractions: At the lowest semantic level our method operates on terms. In most previous work a term can be considered as being synonymous with a word (following the removal of stopwords, as explained previously). Herein we adopted this approach too, henceforth referred to as the unigram approach, but also explored two alternatives. Both of the alternatives are bigram based i.e. they consider pairs of consecutive words in pre-processed sentences. The first one is the ordered bigram, which takes into account the ordering of the words in a pair, and the second the unordered bigram, which discards the ordering information. Thus, the sentence fragment:

"...some pain on some days..."

would result in unigram terms (term frequencies are stated in the corresponding brackets):

‘Some’ (2), ‘pain’ (1), ‘days’ (1),

in ordered bigram terms:

‘some pain’ (1), ‘pain some’ (1), ‘some days’ (1),

and in ordered bigram terms (using alphabetical ordering of words within pairs, without loss of generality):

‘pain some’ (2), ‘day some’ (1).

The intuition behind the use of bigrams in the specific context of interest in the present paper lies in the observation that words indicative of emotional distress (e.g. ‘worry’, ‘concerned’, ‘discomfort’, etc.) are often modified in a manner that provides further evidence (e.g. ‘a lot of worry’, ‘rather concerned’, ‘frequent discomfort’, etc.).
2) Term set representations: The representation of an entire utterance is based on the statistics of occurrences of terms (the meaning of which can be any of the three discussed previously) within it. We evaluated a number of different approaches of doing this and herein we present two which overall produced the most promising results. These are variations on well known representations in the existing literature, adapted to the problem at hand. Specifically, we used:

- binary bag of terms [16], and
- tf-idf (term frequency-inverse document frequency) score [18].

In the binary bag of terms representation used in this work, each entry \( x_i \) in a feature vector \( x_{bow} = [x_1, x_2, \ldots, x_n] \) corresponds to a particular word and is coded as either present in a particular utterance (value 1) or absent from it (value 0), where \( x_{bow} \in \mathbb{R}^n \) and \( n \) is the dictionary size over which the feature vector is constructed (we will discuss vocabulary construction in more detail shortly).

The tf-idf representation \( x_{tfidf} \in \mathbb{R}^n \) too has entries which correspond to different vocabulary terms, but with values which measure the discriminative importance of a particular word in an utterance. Specifically, a particular entry is equal to the frequency of the corresponding term (‘term frequency’) weighted by the inverse number of utterances in the training corpus which also contain the term (‘document frequency’). This representation has been used successfully in a variety of applications, from text mining [18] to visual object recognition and retrieval [3].

a) Vocabulary construction: We constructed the dictionary over which utterance representations (i.e. the corresponding feature vectors) are formed using the widely used approach whereby the most frequent words are included, after the removal of so-called stopwords (e.g. ‘and’, ‘of’, ‘in’, etc.). In particular we determine the vocabulary size by including in it the number of the most frequent words which cover at least 95% of the entire training corpus (by word count). This resulted in a vocabulary consisting of the 300 most frequent words.

D. Classification

Our choice of classifiers used in this work was guided by several different goals. Our primary aim was to assess the feasibility of machine driven prediction at the task at hand. Hence, we selected classifiers which have an established record of good performance across a range of different domains. In addition, we sought classifiers which are well understood but differ in terms of the shape of classification boundaries they can learn. This was done in an effort to facilitate better understanding of data by means of comparative analysis, and therefore provide guidance for future work. Following these criteria, we adopted the use of naïve Bayes and logistic regression based classifiers, support vector machines [19], and boosted ensemble decision trees. For the SVM classifier we used the Gaussian kernel with the scale set to \( 0.25 \sqrt{n_p} \) where \( n_p \) is the number of predictors.

IV. RESULTS AND DISCUSSION

Consultations lasted between 2m 13s and 26m 19s with the average of 8m 46s and the standard deviation of 4m 55s, with 2.2 consultations per patient on average. The split in leading the consultations between the two radiotherapists was 109 and 92. Inter-rater reliability analysis of the frequency and sequencing of codes, based on around 5% of consultations, indicated substantial agreement between the two researchers implementing VR-CoDES (\( \kappa = 0.67 \), 95% CI=0.58–0.75; Spearman’s \( \rho = 0.98, p < 0.001 \)). There were 888 cues and concerns (on average 4.4 per consultation, with the standard deviation of 0.35), and 888 corresponding responses. Considering the rather specific nature of our textual corpus, we first went on to examine if the overall statistical properties of word occurrences within it follow the same type of pattern as in more commonly used, general texts. As the graph of unigram term frequencies plotted against the unigram rank in Figure 1 demonstrates that the power law (or Zipfian) distribution captures the behaviour well. Interestingly, we found the same behaviour with the use of both ordered and unordered bigrams, not only in its functional form but also the distribution parameter. At the same time it is important to note that the value of this parameter which governs the rate of frequency drop-off as the term rank is increased, is rather different than that reported in previous work on short text analysis e.g. Twitter messages [8]. The slower drop-off in our corpus resonates with our observation that medical consultations are rather constrained semantically by their very nature.

The key classification results using different types of terms and classifiers is summarized by the graph in Figure 2. The best performance was achieved using our unordered bigram representation, with very little difference across classifiers. The greater variability observed by different term types suggests strongly that the unordered bigram is inherently a superior one. Using a Gaussian kernel support vector machine, the precision, recall, AUC, F-score, and \( \kappa \) values were 0.93, 0.86, 0.75, 0.72, and 0.45 respectively.

The performance of the tf-idf utterance representation is summarized by the graph in Figure 3. It is readily apparent that the classification performance achieved is outstanding and improved over the simple bag of terms. Moreover, there are some noteworthy qualitative differences. In particular, note that with the use of tf-idf weighting, the dependence
of classification performance across different term types was much reduced, with the ordered bigram now being on par with the unordered bigram, and the unigram approach was significantly inferior only with the use of naïve Bayes.

V. SUMMARY AND CONCLUSIONS

In this paper we described the first attempt at fully automatic labelling of textual transcripts of HCP-patient consultations using the VR-CoDES coding framework. We examined different low and high level representations of patient utterances, and evaluated the effectiveness of a number of well-known classifiers in labelling utterances as either containing an expression of emotional distress explicitly (a ‘concern’ in the VR-CoDES vernacular) or implicitly (a ‘cue’), or neither. The extremely high accuracies (≈ 94%) of our algorithms achieved on a data corpus comprising 200 consultations between radiotherapists and adult female breast cancer patients strongly motivate further work in this area. Our current efforts are focused on the labelling at the deeper cancer patients strongly motivate further work in this area.

REFERENCES