

Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <http://orca.cf.ac.uk/124582/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Spasic, Irena, Owen, David, Smith, Andrew and Button, Kate 2019. KLOSURE: Closing in on open-ended patient questionnaires with text mining. *Journal of Biomedical Semantics* file

Publishers page:

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



KLOSURE: Closing in on open-ended patient questionnaires with text mining

Irena Spasić^{1§}, David Owen¹, Andrew Smith², Kate Button²

¹School of Computer Science & Informatics, Cardiff University, Cardiff, UK

²School of Psychology, Cardiff University, Cardiff, UK

²School of Healthcare Sciences, Cardiff University, Cardiff, UK

[§]Corresponding author

Email addresses:

IS: SpasicI@cardiff.ac.uk

DO: OwenDW1@cardiff.ac.uk

AS: SmithAP@cardiff.ac.uk

BK: ButtonK@cardiff.ac.uk

ABSTRACT

Background: Knee injury and Osteoarthritis Outcome Score (KOOS) is an instrument used to quantify patients' perceptions about their knee condition and associated problems. It is administered as a 42-item closed-ended questionnaire in which patients are asked to self-assess five outcomes: pain, other symptoms, activities of daily living, sport and recreation activities, and quality of life. We developed KLOG as a 10-item open-ended version of the KOOS questionnaire in an attempt to obtain deeper insight into patients' opinions including their unmet needs. However, the open-ended nature of the questionnaire incurs analytical overhead associated with the interpretation of responses. The goal of this study was to automate such analysis. We implemented KLOSURE as a system for mining free-text responses to the KLOG questionnaire. It consists of two subsystems, one concerned with feature extraction and the other one concerned with classification of feature vectors. Feature extraction is performed by a set of four modules whose main functionalities are linguistic pre-processing, sentiment analysis, named entity recognition and lexicon lookup respectively. Outputs produced by each module are combined into feature vectors. The structure of feature vectors will vary across the KLOG questions. Finally, Weka, a machine learning workbench, was used for classification of feature vectors.

Results: The precision of the system varied between 62.8% and 95.3%, whereas the recall varied from 58.3% to 87.6% across the 10 questions. The overall performance in terms of F-measure varied between 59.0% and 91.3% with an average of 74.4% and a standard deviation of 8.8.

Conclusions: We demonstrated the feasibility of mining open-ended patient questionnaires. By automatically mapping free text answers onto a Likert scale, we can effectively measure the progress of rehabilitation over time. In comparison to traditional closed-ended questionnaires, our approach offers much richer information that can be utilised to support clinical decision making. In conclusion, we demonstrated how text mining can be used to combine the benefits of qualitative and quantitative analysis of patient experiences.

Keywords: text mining, natural language processing, text classification, named entity recognition, sentiment analysis, patient reported outcome measure, open-ended questionnaire

BACKGROUND

Musculoskeletal pain is recognised globally as negatively impacting healthy aging and accounts for 21.3% of total years lived with disability [1]. It is associated with frailty, loss of function and

independence during everyday activities and reduced overall physical and mental wellbeing [2]. The knee is one of the most commonly affected joints reportedly affecting 30% of people with joint pain [3]. Knee osteoarthritis is one of the most common conditions and affects 18% of people over the age of 45 in England [4]. With an aging population, this number is set to increase, placing greater burden on health resources [4], adding to waiting lists and causing delays in receiving appropriate care [5]. Therefore, self-management treatment approaches, which equip patients with the skills to manage their health condition, are essential. When away from the secondary care setting, patient reported outcome measures (PROMs) can be used to monitor their health status remotely.

PROMs are standardised, validated questionnaires completed by patients in an attempt to measure their own perceptions of their health conditions. Patient responses are converted into a numerical score, which can be used to monitor patient progress over time and plan treatment accordingly. The Knee injuries and Osteoarthritis Outcome Score (KOOS) [6] is one of the most widely used PROMs for assessing patients' opinions about their knee condition. It is administered as a 42-item closed-ended questionnaire in which patients are asked to assess five outcomes: pain, other symptoms, activities of daily living, sport and recreation activities, and quality of life. The resulting scores on a scale of 0-100 can help both patients and clinicians to monitor the progress of knee rehabilitation. However, KOOS does not capture details surrounding particular patient circumstances [7]. By forcing the respondents to choose from ready-made options, closed-ended questions restrict freedom and spontaneity of responses and as such they are unlikely to tap into the full range of positive and negative expressions of patients [8]. Alternatively, modifying KOOS into an open-ended questionnaire has got a great potential to inform clinicians about patients' opinions including their unmet needs, but this incurs analytical overhead associated with the interpretation of responses.

Unfortunately, patient experience questionnaires remain largely quantitative in nature [9] despite the findings that they tend to overestimate patient satisfaction [10] and that qualitative analysis tends to uncover more actionable information [11]. This can be explained partly by the lack of knowledge on how best to collect and present patient's responses to the stakeholders [10]. From a practical point of view, the cost of qualitative analysis in terms of time and labour play a major factor in its prevalence or the scale of such studies. Poorer statistical significance has often been used as an excuse to dismiss valuable information that can be provided by qualitative research [12]. In light of these issues, text mining (TM), which aims to discover patterns, relationships and trends within text documents, has found a great many biomedical applications [13]. In particular, it is increasingly used to analyse patient experiences, i.e. their thoughts, feelings and behaviours, expressed in their own words [14]. Therefore, TM may support scalability of qualitative analyses of open-ended questionnaires.

To date, most TM approaches used to support the analyses of open-ended questionnaires focused on aggregating all responses across a surveyed population, e.g. consumers [15], students [16, 17], patients [18], etc. Early techniques used to process open-ended questionnaires included rule-based text classification approaches [15]. The proliferation of user-generated data on the Web encouraged the use of data mining, e.g. clustering [16] and association rule mining [19]. The rising popularity of supervised machine learning approaches paved the way to classifying individual responses. In terms of the overall aim and techniques applied, the work on screening patients for posttraumatic stress disorder is the closest to our own [20]. They used an open-ended questionnaire to elicit self-narratives from participants who experienced a traumatic event. They used supervised machine learning to implement a binary classifier with an aim to automatically diagnose a participant, based on their overall response, as having or not having posttraumatic stress disorder. Our approach goes a step further by classifying a response to each open-ended question separately against multiple classes.

METHODS

The aim of this study was to automate measurement of health outcomes from patients' free-text responses to open-ended questions. Addressing this aim required us to: (1) develop an open-ended questionnaire, (2) collect responses to the questionnaire, (3) analyse the responses manually to establish the ground truth, (4) develop text mining methods to analyse the responses automatically, and (5) evaluate the performance of the text mining methods. The following sections describe these steps in more detail.

Open-ended questionnaire

Questionnaire development

We developed an open-ended questionnaire to capture patients' opinions about all aspects relevant to assessing the management of their knee condition while minimising the number and complexity of questions. We designed KLOG (contracted from 'knee log') as an open-ended version of the KOOS questionnaire [6]. Like KOOS, KLOG was designed to elicit responses that can be used to assess five outcomes: pain, other symptoms, activities of daily living, sport and recreation, and quality of life. The open-ended nature of the questionnaire enabled us to reduce the number of questions from 42 in KOOS to only 10 in KLOG. For example, KOOS contains 9 closed-ended questions related to pain: (P1) How often do you experience knee pain? What amount of knee pain have you experienced the last week during the following activities? (P2) twisting/pivoting on your knee, (P3) straightening knee fully, (P4) bending knee fully, (P5) walking on flat surface, (P6) going up or down stairs, (P7) at night while in bed, (P8) sitting or lying, (P9) standing upright. KLOG compresses them into a single open-ended question: "*Can*

you describe any knee pain you have experienced over the past week?" To illustrate a greater coverage of the open-ended question, we provide a sample of answers:

1. *<P1>Constant</P1> pain whether in <P8>sitting</P8> or <P9>standing</P9>.*
2. *<P1>Occasional</P1> sharp pain as well as ache, especially <NA>walking downhill</NA> or on <NA>uneven ground</NA>.*
3. *No pain when <P8>sitting</P8>, some pain when <P6>walking upstairs</P6> and <NA>walking long distances</NA>.*
4. *<P1>Occasional</P1> sharp pain especially when <P6>going up steps</P6>.*
5. *Some general pain when <NA>exercising</NA> and quite painful in the joint for <P1>the last 2 days</P1>.*
6. *I have severe aching and bad pain when I <NA>get up from sitting</NA> and <P5>general walking</P5>.*
7. *<P1>Occasionally</P1> the knee joint aches together with momentary sudden stabbing pain in various parts of the knee. Also, pain if I am <P8>sat</P8> with the knee <P4>bent</P4> such as <P8>sat</P8> in an office chair.*
8. *<NA>Exercise</NA> induced occasional medial and lateral knee joint discomfort knee. Plus some <P7>nocturnal</P7> discomfort.*

We used XML tags to relate free-text answers to the corresponding questions in KOOS as shown in the above answers. We can see that when answering a more generic open-ended KLOG question, patients do provide information related to the corresponding closed-ended questions in KOOS, but they also provide a much richer account of circumstances surrounding their experience of pain. We annotated such references using the NA tag to indicate that the corresponding closed-ended question is not available in KOOS.

To pre-test the KLOG questionnaire, we administered it to a small number of knee patients (16 in total). Participants were asked to complete the questionnaire on a weekly basis over the course of 4 weeks. We obtained a total of 30 responses. The size of the dataset was 4,625 words, which is equivalent to 10.3 A4 pages using Arial 12pt and single spacing. We analysed the data qualitatively to identify potential issues with the questionnaire. Based on the results, the original questions were re-phrased and re-ordered where appropriate. For example, the original question *"Can you describe any knee symptoms you have experienced over the past week?"* was re-phrased to *"Can you describe any knee symptoms other than stiffness and pain you have experienced over the past week?"* and moved after the questions about stiffness and pain to prevent overlapping answers.

Online questionnaire administration

To collect answers remotely, we implemented a web site using a responsive design for user–friendly access on a range of Internet–enabled devices. In the online version of the KLOG questionnaire, users were asked to answer a series of 10 questions in a fixed order:

- Q1. What knee condition are you currently receiving treatment for?
- Q2. What treatment are you currently using for your knee?
- Q3. Have there been any changes to your knee condition over the past week?
- Q4. How confident do you feel about looking after your knee?
- Q5. Can you describe any knee stiffness you have experienced over the past week?
- Q6. Can you describe any knee pain you have experienced over the past week?
- Q7. Can you describe any knee symptoms other than stiffness and pain you have experienced over the past week?
- Q8. Can you describe if your knee condition limited your ability to carry out your day to day tasks over the past week?
- Q9. Can you describe if your knee condition limited your ability to carry out your work, hobbies or exercise over the past week?
- Q10. Do you have any other comments about your knee condition or the treatment you are receiving?

All questions but the last were mandatory. A progress bar was used to visualize the progression through the questionnaire in an attempt to increase completion rates. The access to the questionnaire was gated through a consent form and participant information sheet. All data collected online was stored securely in a back–end database.

Questionnaire response analysis

To automatically process patients' responses to the KLOG questionnaire, we implemented a text mining system, KLOSURE (contracted from 'KLOG measure'). The analysis is performed for each question independently. To interpret answers to questions about the knee condition (Q1) and its treatment (Q2), the system aims to recognise relevant named entities (NEs). For example, to identify what condition a patient is treated for, the system looks for mentions of *medical conditions* (NE1) or *surgical procedures* (NE2) as well as mentions of affected *anatomical structures* (NE3), e.g.

I <NE1>completely ruptured</NE1> my <NE3>ACL</NE3> and had <NE3>ACL</NE3> <NE2>reconstruction surgery</NE2> using a <NE2>hamstring graft</NE2>.

In addition to finding NEs, the system maps them onto the Unified Medical Language System (UMLS) [21], the largest curated collection of biomedical entities. Such mapping normalises representation of NEs by associating all synonyms with a unique identifier. This fact enables text data to be managed, searched and processed based on their meaning. In KLOSURE, named entity recognition (NER) is used not only to provide the final output in relation to questions Q1–Q2, but also to extract NEs as auxiliary features in relation to the remaining questions. The same NER approach is used in both cases. Therefore, to simplify description of the system, we hereby focus on processing answers to questions Q3–Q10 as the main functionality of the system.

KLOSURE aims to classify free-text answers to questions Q3–Q10 on a 3-point Likert scale. This task is known as ordinal classification, where the rating of a data item is estimated on a fixed, discrete rating scale. Ordinal classification can be cast as a supervised learning task, whose goal is to induce a classification model from the training data annotated with class labels. A number of supervised learning methods can be used for this purpose [22], e.g. naïve Bayes classifier, support vector machines, nearest neighbour, decision tree learning, neural networks, etc. These classification methods commonly operate under the assumption that the classes are not ordered, but can be easily adapted to make use of ordering information in class attributes to improve their performance over the naïve approach, which treats the class values as an unordered set [23]. According to the "no free lunch" theorem, any two learning algorithms are equivalent when their performance is averaged across all possible problems [24]. This theorem suggests that the choice of an appropriate algorithm should be based on its performance for the particular problem at hand and the properties of data that characterise the problem. Our choice was based on the results of cross-validation, which will be described later. First, let us discuss feature extraction, which is an important factor in terms of efficiency and classification performance.

Feature extraction

The first problem associated with text classification is high dimensionality of the feature space. The bag-of-words representation suffers from the curse of dimensionality (or Hughes effect), where, given a fixed size of the training dataset, the predictive power of a machine learning algorithm reduces as the dimensionality increases [25]. When the ratio of the number of training instances to the number of features is low, overfitting is likely to occur. Therefore, feature selection is essential to reduce overfitting. Several feature selection methods have been proposed specifically for the problem of ordinal text classification [26]. However, these methods are based on statistical measures of informativeness such as information gain or inverse document frequency, which means that their use requires a relatively large sample size.

Alternatively, when training data are sparse, prior knowledge (i.e. any form of knowledge that may be incorporated prior to training) can be used to extract features based on their general properties without measuring their value on the training data [27]. These properties may include types of features, relationships between features, indications of feature relevance, etc. [28]. One way of incorporating prior knowledge into otherwise agnostic machine learning methods is to pre-process the training data [29]. This step consists of selecting, cleaning and transforming the original data in order to reduce noise and complexity, and increase transparency of the classification model. For example, incorporating prior knowledge into a feature vector document representation can improve text classification accuracy [30]. However, knowledge engineering itself can be fairly time and labour consuming. Ideally, prior knowledge should be re-used from existing sources or otherwise incur as little human intervention as possible. In accordance with these principles, we re-used knowledge readily available from the UMLS [21], the largest curated collection of biomedical entities, which currently covers over 3,640,132 entities and 11,757,373 names that refer to these entities. All entities are organised into a network of 133 semantic types.

Lexico-semantic features

To utilise the knowledge encoded in the UMLS for text mining purposes, we need to map its content onto free text. This is practically achieved by recognising NEs, i.e. names used to differentiate between entities of the same semantic type (e.g. *stiffness* and *pain* are names used to refer to specific *symptoms*), followed by normalising the representation of their meaning (e.g. *swelling* is also known as *edema* or *oedema*, all of which are associated with the same identifier). We used MetaMap, a software tool for recognising UMLS entities in biomedical text [31], to automatically identify NEs of relevant semantic types. In effect, the use of MetaMap allowed us to transform "surface" lexical features into "deep" semantic ones. For instance, when classifying answers to question Q9 about work, hobbies and exercise, we used MetaMap to identify NEs that represent instances of *daily or recreational activity* (DORA), *occupation or discipline* (OCDI), *occupational activity* (OCAC) and *professional or occupational group* (PROG), e.g.

Not been able to carry out my normal <OCDI>job</OCDI> at <OCAC>work</OCAC> due to my injury as I am a commercial <PROG>vehicle mechanic</PROG>. Limited <DORA>exercise</DORA> and <DORA>hobbies</DORA> I haven't been able to play <DORA>squash</DORA> due to the injury and <DORA>golf</DORA> as I can't twist my knee fully yet.

References to these NEs can then be aggregated into a single semantic feature that encompasses work, hobbies and exercise. If needed, one can still differentiate between NEs of the same type. For example, when classifying answers to question Q7 in terms of overall severity, it may be useful to differentiate between various symptoms, e.g. *swelling*, *bruising*, *giving way*, *popping*, etc. By using MetaMap to map

text onto the UMLS, we can aggregate multiple lexical features (e.g. *swelling*, *edema*, *oedema*, *dropsy*) into a single semantic feature, which represents the corresponding concept (e.g. *an abnormal accumulation of fluid beneath the skin or in one or more cavities of the body*). In summary, MetaMap can be used to extract semantic features at two levels of granularity (entity vs. type). In the context of machine learning, both extremes may lead to overfitting, one because of sparsity and the other because of overgeneralisation. Therefore, it may be useful to group related entities to tackle data sparsity without overgeneralising them. Consider, for instance, question Q6 about pain, where answers need to be classified on a severity scale: 1 (none) – 2 (some) – 3 (severe). Our training data contains references to *mild pain*, *severe pain*, *sharp pain*, *stabbing pain*, *intermittent pain* and *constant pain*, which are modelled as distinct entities in the UMLS. Other than grouping these entities together within the same semantic type, the UMLS does not offer any other information about their properties or relationships among them that could be useful features for ordinal text classification, e.g. that *mild pain* < *severe pain* in terms of severity or that *sharp pain* is severe. In addition, when patients discuss their pain, they often use descriptive phrases (e.g. *slight pain* or *bad pain*) rather than established terms (e.g. *mild pain* or *severe pain*). Such laymen's terms are not represented in the UMLS and, therefore, cannot be identified in text using MetaMap.

For the purposes of ordinal text classification, it would still be useful to have some prior knowledge about such laymen's terms, e.g. that *slight pain* < *bad pain* on a severity scale. Given that questions in our study are focusing on specific aspects of self-care, we can limit hand-crafting prior knowledge to these aspects (e.g. pain severity). We can use approaches from corpus linguistics to quickly extract relevant lexical features, and model relevant knowledge around them. For example, to systematically collect adjectives that are commonly used to modify the noun *pain*, we automatically extracted its collocations from a general corpus such as the British National Corpus (BNC) [32]. A search for collocated adjectives, where the strength of collocation was measured by mutual information, retrieved 56 items (see Figure 1). Processing this list did not require extensive manual intervention to remove adjectives that are not applicable in the context of knee pain (e.g. *abdominal pain* or *sweet pain*) and classify the 31 remaining ones in terms of severity (e.g. *overwhelming pain* is more severe than *dull pain*). As we were only interested in laymen terminology, the manual curation did not require medical expertise.

In this manner, we efficiently supplemented prior knowledge available from the UMLS to support extraction of relevant semantic features from raw text. Note that a relatively small scale of the system in term of features and their values did not warrant automation of this approach. However, for the proposed approach to be routinely used on a large scale, a variety of text mining solutions based on distributional semantics (e.g. [33, 34]) can be used to infer relationships between different collocates automatically.

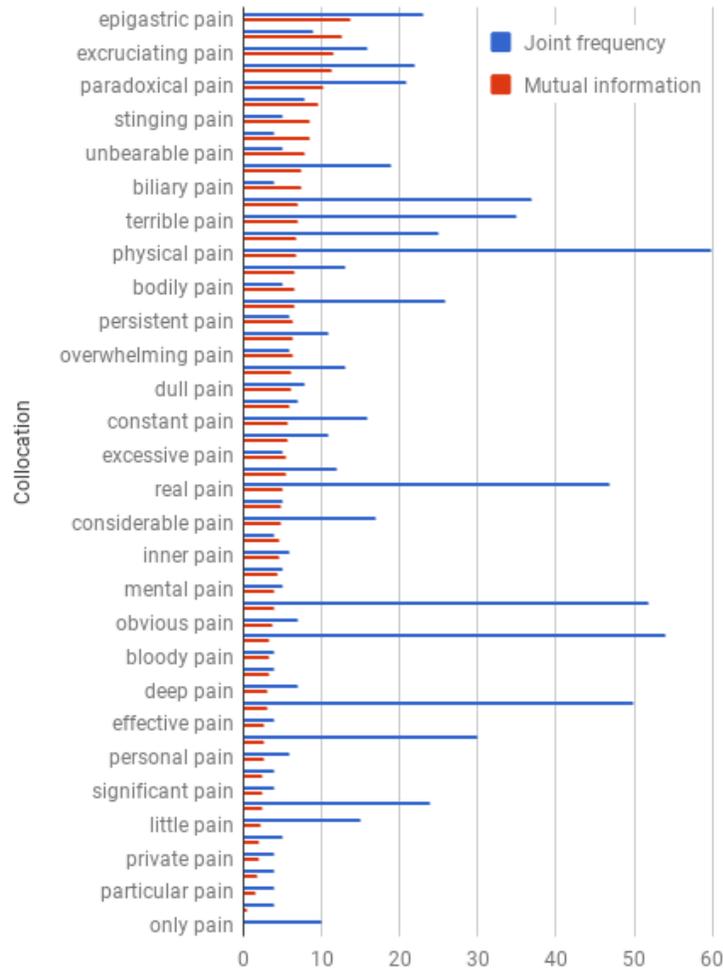


Figure 1: A sample of adjectives collocated with the word *pain*

Negation

Extraction of semantic features would not be complete without considering their context in terms of polar opposition, i.e. affirmation vs. negation [35]. Negated statements contain an additional layer of meaning and morpho-syntactic structure, which incurs an overhead in semantic processing [36]. Naturally, the processing complexity of negation poses considerable challenges in automatically detecting its scope and focus [37]. Some types of explicitly asserted negation can be recognised by using automatically extracted syntactic dependencies. For example, from the syntactic dependency parse shown in Figure 2, it can be automatically inferred that both mentions of the word *pain* are negated.

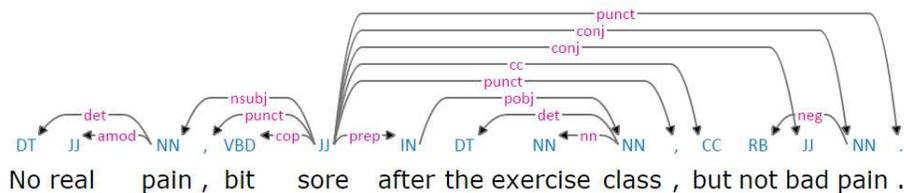


Figure 2: Two negated mentions of the word *pain*

On the other hand, implicit negation, whose source is pragmatic (presupposition or implicature), is not immediately accessible from the syntax and, therefore, is much harder to detect automatically. For example, the verb *to ease* is neither syntactically nor morphologically negative, but it effectively negates *pain* in the sentence given in Figure 3.

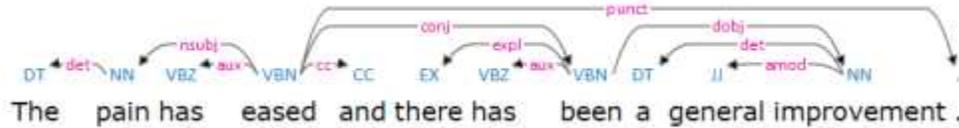


Figure 3: The role of verb *to ease* in effectively negating a nominal subject (nsubj)

In this study, however, we deal only with explicit syntactic negation. We used a parser distributed as part of the Stanford CoreNLP library [38] to extract universal syntactic dependencies, which provide straightforward descriptions of grammatical relations between the words [39]. Specifically, these dependencies were used to identify negated words.

Sentiment polarity

In relation to questions Q5–Q7, we need to classify the severity of symptoms on a 3–point Likert scale: 1 (none) – 2 (some) – 3 (severe). For this, we need to analyse the context in which the symptoms are mentioned. Some of the features discussed thus far can be used for this purpose, e.g. negation and adjectival modifiers. For example, the two negated mentions of *pain* in Figure 2 can be used to classify its severity as 1 (none). Similarly, the collocated adjective *intermittent* can be used to classify the severity of pain in "*Intermittent pain around the joint.*" as 2 (some), whereas the use of adjective *constant* in "*Constant pain whether sitting or standing.*" can be used to classify the pain level as 3 (severe). Combinations of different features make the analysis more complex. Consider, for example, the negated mention of *constant pain* given in Figure 4. The analysis would require us to traverse the syntactic dependency tree to infer that the negation of the verb *to be* applies to the mention of the noun *pain*.

Alternatively, the positive sentiment of the given sentence can be used to classify the severity of pain as either 1 (none) or 2 (some), depending on the strength of the sentiment. Similarly, the negative sentiment expressed in "*My pain was unbearable and I was screaming, crying, and regretting how it happened.*" can be used to classify the pain as 3 (severe). In principle, examples of non–asserted negation given in Figure 3 can also be handled using sentiment polarity. In our system, we used Stanford CoreNLP's sentiment annotator [40], which uses a deep neural network approach to build up sentiment representation of a sentence on top of its grammatical structure.

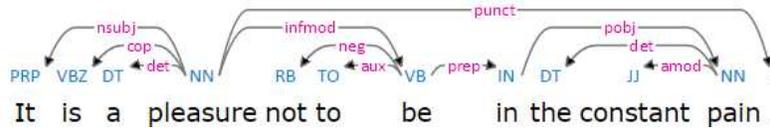


Figure 4: A negated mention of the word *pain*

System summary

The KLOSURE system consists of two subsystems, one concerned with feature extraction and the other one concerned with classification of feature vectors. Table 1 describes the conceptual architecture of the feature extraction subsystem. Module names describe their main functionality. Specific tools used to support these functionalities are referenced in the second column. Resources used by these tools constitute the parameters of the system. Any potential improvements of the system performance will be largely confined to this area. Outputs produced by each module are combined into feature vectors. The structure of feature vectors will vary across the KLOG questions. Finally, Weka [41], a machine learning workbench, was used for classification of feature vectors. Classification is performed against a scheme described in Table 2. Specific classification methods are discussed in the following section in the context of their performance.

Table 1. Feature extraction subsystem.

Module	Software	Resources	Output
linguistic pre-processing	Stanford Core NLP [38]	language model	POS tags, dependencies
sentiment analysis	Stanford Core NLP [40]	sentiment model	sentiment polarities
named entity recognition	MetaMap [31]	UMLS [21]	named entities, semantic types
lexicon lookup	N/A	24 lexicons	matched items

Table 2. Definition of ordinal classes.

Question	1	2	3
Q3	worse	same	better
Q4	no	reasonably	fully
Q5–Q7	none	some	severe
Q8–Q9	not at all	somewhat	a lot
Q10	negative	neutral	positive

RESULTS

Data

To collect the data, we approached individuals with knee conditions in secondary care and relevant online communities to complete the KLOG questionnaire. We collected a total of 55 responses. The size of the dataset is 7,985 words, which is equivalent to 17.7 A4 pages using Arial 12pt and single spacing (note that this dataset is distinct from the one discussed in the Methods section). To establish the ground truth for evaluation of the KLOSURE system, all data were annotated manually. For questions Q1–Q2, annotators were asked to identify all mentions of relevant NEs and categorise them according to their types. For every mention, annotators were asked to identify the maximal extent of the string that represents an entity. Nested mentions were not annotated. The annotation task for questions Q3–Q10 was framed as ordinal text classification on a 3–point Likert scale.

Every response was annotated independently by two experts. Inter–annotator agreement (IAA) for annotations of NEs related to questions Q1–Q2 was calculated using Cohen's kappa coefficient [42], whose values were 0.688 and 0.806 respectively. Cohen's kappa coefficient treats all disagreements equally, which is not suitable when the annotation categories are ordered as they indeed are in relation to questions Q3–Q10. In such case, it is preferable to use weighted kappa coefficient, which accounts for the degree of disagreement [43]. Figure 5 provides the values of the Cohen's kappa coefficient for questions Q3–Q10. The unweighted IAA across all 10 questions ranged from moderate (0.399) to very good (0.901). At average kappa value of 0.693 and standard deviation of 0.132, the overall IAA was found to be good [44]. The ground truth was created by the third annotator who independently resolved all disagreements.

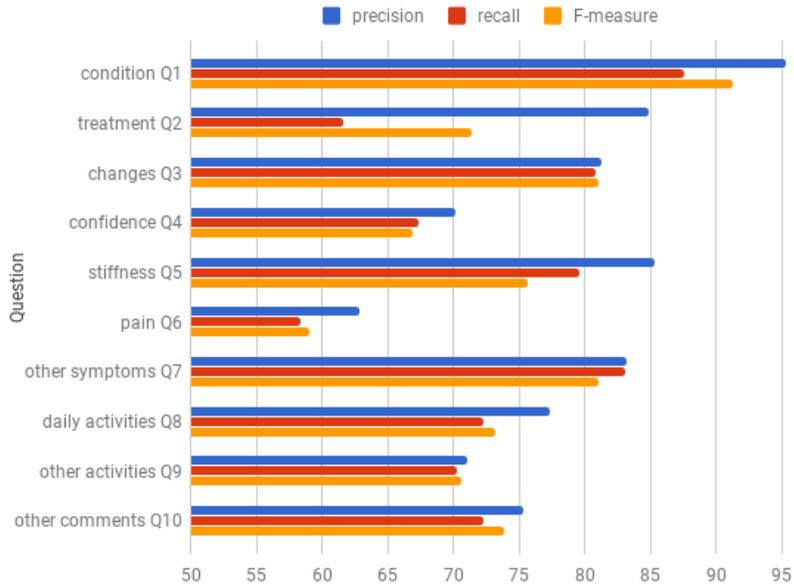


Figure 5: Inter-annotator agreement for questions Q3–Q10

Evaluation

Most text mining tasks can be viewed as classification problems in which, given an instance, the system predicts its class label. For instance, a NER system in effect labels a phrase as a NE of a given type (or class). Various measures can be used to evaluate classification performance based on a confusion matrix, which contains information about actual (or known) labels and those predicted automatically by the system. Frequently used measures are that of precision (P) and recall (R), which are calculated as the following ratios [45, 46]: $P = TP / (TP + FP)$ and $R = TP / (TP + FN)$, where TP, FP and FN are the numbers of true positives, false positives and false negatives respectively. The given formulas are applied to calculate precision and recall for each class separately. The overall precision and recall values are calculated using weighted average, which is simply a sum of class-specific values weighted according to the number of instances within a particular class label. Systems will often be compared on how well they balance precision and recall. For this purpose, F-measure is calculated as their harmonic mean. We used these measures to evaluate the performance of the KLOSURE system.

The classification subsystem consists of eight modules – one classifier per question Q3–Q10. For question Q10, we re-used an existing open-source sentiment analysis tool [40]. To implement the remaining seven classifiers, we used Weka [41], a software workbench that incorporates a standard range of supervised learning methods. To systematically measure the predictive ability of a classification model for each question, we performed 10-fold cross-validation experiments. Table 3 provides the results achieved by the best performing methods. The evaluation results are also visualised in Figure 6.

Table 3. Performance of the KLOSURE system.

Question	Topic	Classes	Method	Features	P (%)	R (%)	F (%)
Q1	condition	3	MetaMap	N/A	95.3	87.6	91.3
Q2	treatment	4	MetaMap	N/A	84.9	61.6	71.4
Q3	changes	3	naive Bayes	8	81.3	80.8	81.0
Q4	confidence	3	best-first decision tree	8	70.1	67.3	66.9
Q5	stiffness	3	reduced error pruning tree	8	85.3	79.6	75.6
Q6	pain	3	complement naive Bayes	10	62.8	58.3	59.0
Q7	other symptoms	3	naive Bayes	5	83.2	83.0	81.0
Q8	daily activities	3	J48 pruned tree	14	77.3	72.3	73.2
Q9	other activities	3	random forest	14	71.0	70.2	70.6
Q10	other comments	3	Stanford Core NLP	N/A	75.3	72.3	73.8

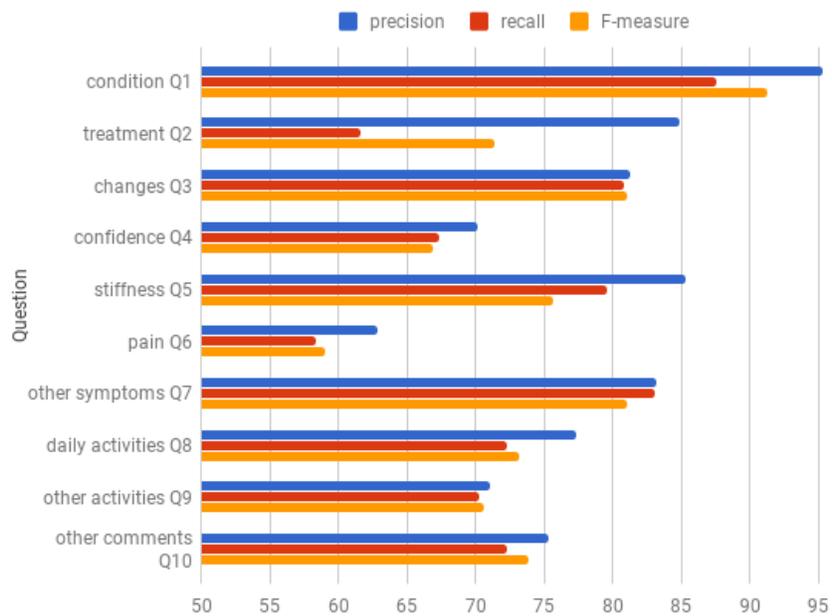


Figure 6: Evaluation results

Discussion

The overall performance in terms of F-measure varied between 59.0% and 91.3% with an average of 74.4% and a standard deviation of 8.8. Given the subjective nature of pain, it is not surprising that determining the severity of pain (Q6) proved to be the most difficult classification task. Even commonly used pain rating scales are plagued with issues related to sensitivity, reproducibility, failure rate and

interpretation of scores [47]. On the other hand, the factual nature of question Q1 about a diagnosed medical condition was associated with the best performance.

Similar performance would be expected in relation to question Q2 about treatments. However, recall was found to be substantially lower. The analysis of false negatives revealed two underlying causes. The first one is related to layman's terms used by patients, e.g. *physio* and *cardio* failed to be recognised as *physiotherapy* and *cardiovascular exercises* respectively. The other one is related to under-representation of knee rehabilitation specific concepts in the UMLS, mostly specific exercises mentioned as part of the physiotherapy, e.g. *hamstring curls*. The TRAK ontology defines concepts relevant for the standard care for the rehabilitation of knee condition including a wide range of exercises [48] and can be used to supplement resources used to support NER.

Similarly, other classifiers could benefit from adding or replacing the resources used (see Table 1). For example, the sentiment analysis module uses a classification model trained on a corpus of movie reviews. Re-training the model on a corpus of patient narratives is expected to improve the performance not only in relation to question Q10 where it provides the final output, but also in relation to questions Q3–Q9 where it is used to extract sentiment-related features.

The lack of explainability has been identified as one of the major obstacles preventing widespread use of machine learning in clinical practice. We can observe that the best performing supervised classifiers tend to be of two varieties: naive Bayes learning and decision trees. The latter may improve the acceptability of the proposed text mining solution in clinical practice because they explicitly provide an explanation in the form of easily interpretable decision trees.

CONCLUSION

In this paper, we demonstrated the feasibility of mining open-ended patient questionnaires. By automatically mapping free text answers onto a Likert scale, we can effectively measure the progress of rehabilitation over time. Arguably, the same can be said about closed-ended questionnaires such as KOOS. However, by spotting a negative trend, in addition to basic alert functionality, our approach offers much richer information that can be utilised to support clinical decision making. Actively listening to patients and involving them into their medical care has shown to improve health outcomes and patient satisfaction [49]. The act of writing itself may have therapeutic benefits [50]. In conclusion, we demonstrated how text mining can be used to combine the benefits of qualitative and quantitative analysis of patient experiences.

Having established the feasibility of the text mining approach, our future work will include validation of KLOSURE as a patient-reported outcome measure. This may be achieved by direct comparison of

KLOSURE against KOOS. To streamline the process of data collection via KLOG, we will explore the use of voice-to-text software, which transcribes spoken words into text. In terms of productivity, text can be dictated three times as fast as typing it. State-of-the-art voice-to-text software is highly accurate, which makes it easy to use at all levels of literacy as it will automatically transcribe words that are otherwise difficult to spell. Moreover, voice commands allow for hands-free interaction, which make voice-to-text software accessible to patients with a wide range of disabilities and injuries that restrict interaction with input devices such as touchscreen display, keyboard and mouse. In summary, speed, accuracy, ease of use and accessibility make voice-to-text software a great user-friendly tool for patients to respond to open-ended questions.

ABBREVIATIONS

BNC - British National Corpus

IAA - Inter-annotator agreement

KLOSURE - KLOG measure

KOOS – Knee injury and Osteoarthritis Outcome Score

NE - Named entity

NER - Named entity recognition

P - Precision

PROMs - Patient reported outcome measures

R - Recall

TM - Text mining

TRAK - Taxonomy for Rehabilitation of Knee conditions

UMLS - Unified Medical Language System

DECLARATIONS

Ethics approval and consent to participate

Ethical approval was gained from North West - Lancaster Research Ethics Committee, reference: 16/NW/0513. NHS Research & Development approval was granted from Cardiff and Vale University Health Board, reference: 16/MAY/6491. All participants in this study provided written informed consent.

Availability of data and materials

The dataset collected and analysed in this study is not publicly available due to data privacy. The code for feature extraction is available on demand from the corresponding author. Table 1 provides references to other publicly available software used in this study.

Competing interests

To the best knowledge of the authors, there are no competing interests.

Funding

This study was partly funded by Wellcome Trust via Institutional Strategic Support Fund. Publication costs are funded by Cardiff University.

Authors' contributions

IS conceived the overall study, designed the system and drafted the manuscript. DO implemented the system and coordinated evaluation. AS supervised qualitative analysis. KB lent her clinical expertise to inform system design and interpret the results. All authors read and approved the final manuscript.

Acknowledgments

This study was originally presented at the Healthcare Text Analytics Conference (HealTAC), Manchester, UK in April 2018.

References

1. Vos T, Flaxman A, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2163-96.
2. Briggs AM, Cross MJ, Hoy DG, et al. Musculoskeletal health conditions represent a global threat to healthy aging: A report for the 2015 World Health Organization world report on ageing and health. *The Gerontologist*. 2016;56(Supplement 2):S243-S55.
3. Thiem U, Lamsfuß R, Günther S, et al. Prevalence of self-reported pain, joint complaints and knee or hip complaints in adults aged ≥ 40 years: A cross-sectional survey in Herne, Germany. *PLoS One*. 2013;8(4):e60753.

4. Arthritis Research UK. Osteoarthritis in general practice - Data and Perspectives. The Medical Press; 2013.
5. Briggs T. Getting it right first time: Improving the quality of orthopaedic care within the National Health Service in England. 2012.
6. Roos EM, Roos HP, Lohmander S, Ekdahl C, Beynon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS) - development of a self-administered outcome measure. *Journal of Orthopaedic & Sports Physical Therapy*. 1998;28(2):88-96.
7. Murphy M, Hollinghurst S, Turner K, Salisbury C. Patient and practitioners' views on the most important outcomes arising from primary care consultations: a qualitative study. *BMC Family Practice*. 2015;16:108.
8. Marcinowicz L, Chlabicz S, Grebowski R. Open-ended questions in surveys of patients' satisfaction with family doctors. *Journal of Health Services Research & Policy*. 2007;12(2):86-9.
9. Iversen HH, Bjertnæs ØA, Skudal KE. Patient evaluation of hospital outcomes: an analysis of open-ended comments from extreme clusters in a national survey. *BMJ Open*. 2014;4(5):e004848.
10. Riiskjær E, Kofoed JAP-E. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective. *International Journal for Quality in Health Care*. 2012;24(5):509-16.
11. O'Cathain A, Thomas KJ. "Any other comments?" Open questions on questionnaires - a bane or a bonus to research? *BMC Medical Research Methodology*. 2004;4(25).
12. Shuval K, Harker K, Roudsari B, et al. Is qualitative research second class science? A quantitative longitudinal examination of qualitative research in medical journals. *PLoS One* 2011 Feb 2011;6(1):e16937.
13. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*. 2015;17(1):33-42.
14. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research*. 2016;25(2):86-100.
15. Yamanishi K, Li H. Mining open answers in questionnaire data. *IEEE Intelligent Systems*. 2002; 17(5):58-63.
16. Hirasawa S, Shih FY, Yang WT. Student questionnaire analyses for class management by text mining both in Japanese and in Chinese. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Montreal, Canada, 2007*, p. 398-405.

17. Huang CJ, Tsai PH, Hsu CL, Pan RC. Exploring cognitive difference in instructional outcomes using text mining technology. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, 2006, p. 2116-2120.
18. Ando M, Morita T, O'Connor SJ. Primary concerns of advanced cancer patients identified through the structured life review process: A qualitative study using a text mining technique. Palliative & Supportive Care. 2007; 5(3): 265-271.
19. Chen YL, Weng CH. Mining fuzzy association rules from questionnaire data. Knowledge-Based Systems. 2009; 22(1): 46-56.
20. He Q, Veldkamp BP, de Vries T. Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. Psychiatry Research. 2012; 198(3):441-447.
21. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(Database issue):D267–D70.
22. Caruana R, Alexandru M. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning; Carnegie Mellon University, Pittsburgh, USA2006. p. 161-8.
23. Frank E, Hall M. A simple approach to ordinal classification. In: Flach LDRP, editor. European Conference on Machine Learning; Freiburg, Germany, 2001. p. 145-56.
24. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural Computation. 1996;8(7):1341-90.
25. Hughes GF. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory. 1968;14(1):55-63.
26. Baccianella S, Esuli A, Sebastiani F. Feature selection for ordinal text classification. Neural Computation. 2014;26(3):557-91.
27. Krupka E, Tishby N. Incorporating prior knowledge on features into learning. Proceedings of the 11th International Conference on Artificial Intelligence and Statistics; San Juan, Puerto Rico, 2007. p. 227-34.
28. Guyon I, Saffari A, Dror G, Cawley G. Agnostic learning vs. prior knowledge challenge. Proceedings of the International Joint Conference on Neural Networks; Orlando, Florida, USA, 2007.
29. Yu T, Simoff S, Jan T. VQSVM: A case study for incorporating prior domain knowledge into inductive machine learning. Neurocomputing. 2010;73(13-15):2614-23.
30. Boghrati R, Garten J, Litvinova A, Dehghani M. Incorporating background knowledge into text classification. CogSci; Pasadena, California, USA2015.

31. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. American Medical Informatics Association 2001. p. 17-21.
32. The British National Corpus (version 3). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk>; 2007, Accessed November 20, 2017.
33. Mikolov T, Sutskever I, Kai Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality, in Proceedings of the 26th International Conference on Neural Information Processing Systems; Lake Tahoe, Nevada, USA, 2013, p. 3111-3119.
34. Pennington, J., R. Socher, and C. Manning, Glove: Global vectors for word representation, in Proceedings of the Conference on Empirical Methods in Natural Language Processing; Doha, Qatar, 2017, p. 1532-1543.
35. Horn LR. A natural history of negation. Chicago, USA: The University of Chicago Press; 1989. 637 p.
36. Xiang M, Grove J, Giannakidou A. Semantic and pragmatic processes in the comprehension of negation: An event related potential study of negative polarity sensitivity. *Journal of Neurolinguistics*. 2016;38:71-88.
37. Blanco E, Moldovan D. Some issues on detecting negation from text. Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference; Palm Beach, Florida, USA, 2011.
38. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP natural language processing toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Baltimore, Maryland, USA, 2014. p. 55-60.
39. de Marneffe M-C, Silveira N, Dozat T, et al. Universal Stanford Dependencies: A cross-linguistic typology. Proceedings of the 9th International Conference on Language Resources and Evaluation; Reykjavik, Iceland, 2014.
40. Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; Seattle, Washington, USA, 2013.
41. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The Weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*. 2009;11(1):10-8.
42. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960;20(1):37-46.

43. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*. 1968;70(4):213-20.
44. Altman DG. *Practical statistics for medical research*: Chapman and Hall, London; 1991.
45. Hersh W. Evaluation of biomedical text-mining systems: Lessons learned from information retrieval. *Briefings in Bioinformatics*. 2005 6(4):344-56.
46. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009;45(4):427-37.
47. Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. *Journal of Clinical Nursing*. 2005;14(7):798-804.
48. Button K, van Deursen RW, Soldatova L, Spasić I. TRAK ontology: Defining standard care for the rehabilitation of knee conditions. *Journal of Biomedical Informatics*. 2013;46(4):615-25.
49. Mercieca C, Cassar S, Borg AA. Listening to patients: improving the outpatient service. *International Journal of Health Care Quality Assurance*. 2014;27(1):44-53.
50. Baikie KA, Wilhelm K. Emotional and physical health benefits of expressive writing. *Advances in Psychiatric Treatment*. 2005;11(5):338-46.

Figures

Figure 1: A sample of adjectives collocated with the word pain

Figure 2: Two negated mentions of the word pain

Figure 3: The role of verb to ease in effectively negating a nominal subject (nsubj)

Figure 4: A negated mention of the word pain

Figure 5: Inter-annotator agreement for questions Q3–Q10

Figure 6: Evaluation results