Using topic modeling to infer the emotional state of people living with Parkinson's disease

Abstract

Individuals with Parkinson's disease (PD) often exhibit facial masking (hypomimia), which causes reduced facial expressiveness. This can make it difficult for those who interact with the person to correctly read their emotional state and can lead to problematic social and therapeutic interactions. In this article, we develop a probabilistic model for an assistive device, which can automatically infer the emotional state of a person with PD using the topics that arise during the course of a conversation. We envision that the model can be situated in a device that could monitor the emotional content of the interaction between the caregiver and a person living with PD, providing feedback to the caregiver in order to correct their immediate and perhaps incorrect impressions arising from a reliance on facial expressions. We compare and contrast two approaches: using the Latent Dirichlet Allocation (LDA) generative model as the basis for an unsupervised learning tool, and using a human crafted sentiment analysis tool, the Linguistic Inquiry and Word Count (LIWC). We evaluated both approaches using standard machine learning performance metrics such as precision, recall, and F_1scores. Our performance analysis of the two approaches suggests that LDA is a suitable classifier when the word count in a document is approximately that of the average sentence, i.e., 13 words. In that case the LDA model correctly predicts the interview category 86% of the time and LIWC correctly predicts it 29% of the time. On the other hand, when tested with interviews with an average word count of 303 words, the LDA model correctly predicts the interview category 56% of the time and LIWC, 74% of the time. Advantages and disadvantages of the two approaches are discussed

Keywords: Parkinson's disease, topic modeling, emotional state, LIWC, latent dirichlet allocation

Introduction

Parkinson's disease (PD) is a universal disorder with an incidence ranging from 9.7 to 13.8 per 100,000 population per year (WHO, 2006). In the US, PD follows Alzheimer's disease as the most common neurodegenerative disorder, affecting at least 500,000 Americans and perhaps 500,000 more if we include the undiagnosed and misdiagnosed cases (*National Institute of Health*, 2018). Tremors, muscle rigidity, bradykinesia (slowness of movement), and loss of balance are symptoms which accompany the disease; it is progressive, the symptoms worsening over time. The first three symptoms can occur in the facial, respiratory, and vocal muscles, resulting in diminished control of one's facial and vocal expression which can dissociate one's inner emotional state from the outward facial appearance; this is known as facial masking and is called hypomimia. Because people rely heavily on facial expression in attributing and interpreting other's emotions and motivational states, facial masking can deeply affect the person's ability to communicate which may lead to impaired social interactions and reduced quality of life (Sturkenboom et al., 2013; Takahashi & Tickle-Degnen, 2010). For example, rehabilitation therapists often use a client's verbal and nonverbal behaviour to infer the client's emotional state; if the client is mostly silent or displaying little facial expression, the therapist may infer the client to be more hopeless or apathetic which may not be their true emotional state. In the home and community, desynchronization between a person's emotional state and her external expression can occur during any social situation, which might take place in the home, among family and friends, and at work (Takahashi & Tickle-Degnen, 2010). This may exacerbate feelings of social incapacitation and stigmatization, which leads to reduced quality of life and the vicious cycle of decreasing social engagement (Ma, Saint-Hilaire, Thomas, & Tickle-Degnen, 2016).

Given that facial expressiveness is a problematic channel for communicating emotions and emotional states in people with PD, a more accurate channel might be verbal communication: the words a person uses in their verbal or written speech (DeGroat, Lyons, & Tickle-Degnen, 2006). Since for humans it is very difficult to override the interpretation of information transmitted through facial expression, which happens automatically and instinctively, it would be helpful to have a reliable, automated way of analysing verbal communication that helps detect the valence of the emotion expressed. This automated capability could be implemented in a communication-assistive tool for improving social life. The tool could take the form of a robotic companion or an application that would help people living with PD, their caregivers, and social community by alerting conversation partners to misunderstandings coming from the desynchronization the person with PD experience of emotion and its reflection in the face. This device is meant to improve natural human interaction in the home and community.

For text, detection of emotional content and its valence has been attempted using an automated textual analysis software program called Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Tausczik and Pennebaker (2010) showed that LIWC's categories for positive emotion, negative emotion, anxiety/fear, anger, and sadness/depression were correlated with external raters' judgments, demonstrating they can be used to assess emotional content in text. However, there are several limitations to using the LIWC approach. The basis for LIWC's text analysis is a dictionary which in the latest version (i.e., LIWC2015) consists of approximately 6,400 words, word-stems, and emoticons, i.e., a pictorial representation of human facial expressions used to convey emotion in text (Pennebaker et al., 2015). LIWC's dictionaries are constructed by human scientists according to

evaluation data generated by human raters, rather than learned from the text automatically. This means that LIWC cannot be used with natural languages for which the software has not been modified to accommodate (i.e. for which such dictionaries have not been created). LIWC relies on word recognition, and needs to be periodically updated as language usage evolves. Also, LIWC is not designed for spoken language (Pennebaker et al., 2015), while for the detection of emotional content in a conversation, the ability to work with spoken language is crucial.

In this paper we introduce a novel approach: using the Latent Dirichlet Allocation (LDA) generative model as the basis for an unsupervised learning tool which is trained to extract topic proportions from a collection of text documents (see the *Background* and *Methods* sections for details). When an unseen document is presented to the model, it finds the document's topic proportions and uses them as a set of features. We then use a logistic regression (LR) classifier to associate these features with training data having enjoyable emotional content (text obtained through the prompt: talk about an enjoyable experience) or frustrating emotional content (text obtained through the prompt: talk about a negative emotional experience). We compare our model with the LIWC approach: the word count frequency of five LIWC features associated with emotion is extracted from the text and these are used to train another LR classifier to associate them with the emotion content labels frustrating and enjoyable that have opposing valence.

The paper proceeds in the following way. In the *Background and related work* section, we review the LIWC and LDA approaches. In the *Methods* section, we show how interview transcripts from the *Self-management Rehabilitation and Health-Related Quality of Life in Parkinson's disease* database (Tickle-Degnen, Ellis, Saint-Hilaire, & Wagenaar, 2010) were used as text documents to train and test both models and we

compare the results of two experiments: the first using training and test documents from the entire interview (average word count = 303) and the second using documents which were edited to contain the first 20 seconds of the interview transcripts (average word count = 13). The Results section shows that for longer text, the LDA model correctly predicts the emotion label (frustrating or enjoyable) 56% of the time while the LIWC model 74% of the time. However, in the case of shorter text, the LDA model outperforms the LIWC model. We then discuss advantages and disadvantages of each approach and potential ways of using them to create emotion detecting assistive conversation tools.

Background and related work

Human-curated approach: LIWC

A reliable method for analyzing the emotional content of text is useful in a wide range of scenarios such as opinion mining where it is necessary to detect shifts in customer sentiment as expressed in social media. One approach is to manually label words according to their semantic valence, either positive or negative (Liu, 2010), creating a sentiment lexicon. Generating a reliable sentiment lexicon manually is time-consuming and thus most researchers rely on already-generated lexicons such as LIWC (Hutto & Gilbert, 2014). LIWC was first introduced in 1993 and has been updated three times since; its latest version was released in 2015. As previously indicated, LIWC2015 contains an internal default dictionary that is used to determine the words which should be counted in the documents. The dictionary of approximately 6,400 words is associated with particular domains, such as negative emotion, and these are called *word categories*. There are 41 word categories associated with a psychological category (e.g., affect, biological processes), six personal concern categories (e.g., home, work, leisure), five

informal language markers (e.g., swear words, net-speak), and 12 punctuation categories. When a word in the text is found in the dictionary, all the word categories that it belongs to have their counts incremented (Pennebaker et al., 2015). The reliability of LIWC has been validated internally (e.g., checking whether the more a person uses a word from a LIWC word category in a text, the more the person uses other words from the same category). The external validity of the LIWC categories have been assessed in contexts relevant to daily living and mental and physical health (Tausczik & Pennebaker, 2010).

With regards to emotional expression in PD, Takahashi and Tickle-Degnen (2010), using data from the same database as this study, measured expressive behavior in transcripts of 212 video clips of 106 persons living with PD by using LIWC to count the number of motivation-related words in each transcript. The videos were recordings of interviews in which the participants were asked to discuss an enjoyable or frustrating activity that occurred during the past seven days. The researchers reported that when participants discussed enjoyable activities, they tended to use more words associated with the LIWC positive emotion category compared to when they discussed frustrating activities. Conversely, participants tended to use more words associated with the LIWC negative emotion category when discussing frustrating activities. The research objective of the current study is to determine whether our machine learning model can achieve similar results to LIWC, using the participants' interview transcriptions from the Takahashi and Tickle-Degnen (2010) study and from the Tickle-Degnen et al. (2010) study.

Latent Dirichlet Allocation (LDA)

Generating and maintaining a sentiment lexicon suitable for reliably extracting emotional content and its valence from text is a labor and time-intensive undertaking.

For example, there have been three major releases since LIWC's initial release in 1993, each containing a new dictionary and improved software design, the result of human testing and validation as well as software engineering effort (Pennebaker et al., 2015). To this end, automated approaches to identifying and extracting features from documents which are correlated with emotion valence and intensity have been the subject of active research. We categorize these approaches as *machine learning*, i.e., the fields of study in which computers learn without explicitly being programmed. In contrast to LIWC in which humans have carefully associated words to emotion categories via its dictionary, the challenge for designing a machine learning model is to identify the features contained in the text, i.e. characteristics of the text that can be used to consistently identify distinctive categories, such as enjoyable vs. frustrating emotional content. The goal is to find features such that as words associated with emotion valence change or new ones are introduced, the model's features also adapt.

Such features can be found in the thematic structure of a document. Topic modeling is the detection of the thematic structure of a document collection; it is a classic problem in natural language processing. One of the motivations for research in this area is to find ways to reduce the dimensionality of large collections of text; the goal is to find semantic structures in the text, which can be used to represent its characteristics using a parsimonious amount of information. This lower-dimensional representation can be used, for example, as an efficient way to retrieve the text.

If samples of text were obtained, we hypothesize that a collection of text documents will contain a mixture of topics. The proportions of these topics in a single document could reflect the enjoyable and frustrating topics contained in that document. Thus, the model design goal is to detect thematic, topic information contained in a sufficiently large sample of text (i.e., a document collection) so that when a document

the model has not yet seen is presented, it can identify the proportion of topics contained therein. We then train a *classifier* to associate a large sample of documents whose emotion valence is already known with these topic proportions. Once that is done, we now have created a way to predict the valence of the emotional content (e.g. enjoyable or frustrating) of any document for which we have extracted its topic proportions. For the feature extraction component of our model design, we will draw from the field of topic modeling, using a technique called Latent Dirichlet Allocation (Blei, Eng, & Jordan, 2003).

LDA is built around the intuition that documents exhibit multiple topics (Blei et al., 2003). LDA makes the assumption that only a small set of topics are contained in a document and that they use a small set of words frequently. The result is that words are separated according to meaning and documents can be accurately assigned to topics. LDA is a generative data model which as the name implies describes how the data is generated. This idea is to treat the data as observations that arise from a generative, probabilistic process, one that includes hidden variables, which represent the structure we want to find in the data. For our data, the hidden variables represent the thematic structure (i.e., the topics) that we do not have access to in our documents. Simply put, a generative model describes *how* the data is generated, and *inference* is used to backtrack over the generative model to discover the set of hidden variables which best explains how the data was generated. To express the model as a generative probabilistic process, we start by assuming that there is some number of topics that the document contains and each topic is a distribution over terms (words) in the vocabulary. Every topic contains a probability for every word in the vocabulary and each topic is described by a set of words with different probabilities reflecting their membership in the topic. The LDA generative process can be described as follows:

For each document:

(1) Choose a distribution (i.e., list of topic proportions) over the topics in the document: $P(\Theta_d)$, which is the per-document topic proportion drawn from a Dirichlet distribution. Note that we have a collection of documents and are choosing a distribution for one of the documents in the collection. The eponymous Dirichlet in Latent Dirichlet Allocation is the name of the distribution that can be used to sample from a collection of distributions.

(2) Repeatedly draw a topic from this distribution. Draw a word, w, from the distribution of words for that topic, with the probability: $P(w|Z, \beta_k)$, where Z is the hidden topic assignment and β_k is the topic distribution over all the words in the vocabulary. Note that β_k is a Dirichlet distribution as we have a collection of topics from which we are choosing a distribution over words.

For another document repeat (1) and (2). The above process generates each document on a word by word basis, according to the assumptions made about the document's thematic structure (i.e., topic proportions and word distribution), regardless of word order; this latter characteristic is known as a *bag of words* model. We never get to observe this structure, so it must be inferred by asking: (i) what are the topics that generated these documents? (ii) for each document, what is the distribution over the topics associated with that document? (iii) for each word, which topic generated the word? In other words, we want to infer the topic structure which can be thought of, in probabilistic terms, as computing the posterior distribution of our generative model: $(P_{hv} \mid P_o)$, where P_{hv} is the probability that the document collection has a thematic structure given P_o, the probability of observing the document collection. Operationally, the hidden variables represented by P_{hv} can be computed several ways using a class of

algorithms known as approximate posterior inference. In our model, the LDA algorithm computes both the hidden variables Z (per-word topic assignment) and Θ_d (per-document topic proportion). We hypothesize that the topic proportions are features which are reduced-dimensionality representations of the original documents and preserve essential characteristics such as the valence of the emotional content of the text. Our model uses a machine learning classifier to systematically correlate these features with PD participants' interviews, labelled according to their enjoyable or frustrating emotional content.

Methods

Materials

Input to our model is a document collection of de-identified transcribed interviews collected during a previously conducted randomized control trial called *Self-management Rehabilitation and Health-Related Quality of Life in Parkinson's disease* (Tickle-Degnen et al., 2010). Data for the current study include responses to open-ended questions about daily life events in the recent past that participants had experienced as particularly frustrating or enjoyable. Participants (N = 117) were people in the early to middle stages of PD, with mild unilateral or bilateral symptoms, Hoehn & Yahr stages 1 through 3 (Goetz et al., 2004), were unassisted for walking and communicating, non-depressed, and of normal mental status. Of the 117 participants, 69.8% were male and 30.2% were female with an average age of 65.6; on average, participants were diagnosed with PD seven years prior to the study. At the time of the interview, participants were "on stage" (i.e., they were taking their medication and their medication was working).

Using a mood-manipulation protocol, the researchers examined the participants'

apparent emotional state by asking them to recall two types of experiences: a frustrating one and an enjoyable one that they had during the past seven days. The interviews were videotaped, later transcribed and the response to each prompt was saved in a separate document. The interviews were conducted at the following intervals: at the baseline, after six weeks, and then two months and six months, post-intervention. Participants talked about typical activities with a focus on their social life and interactions.

Since extracting features using LIWC requires at least some words to count in order to correlate with the built-in emotion categories, we created one dataset containing only documents with at least 130 words to be included in our models, resulting in a document collection of 366 positive and negative interviews. Documents contained an average word count of 303 words, with the largest containing 1732 words and the smallest, 131. We also created a dataset of 448 documents containing documents with an average word count of 258 words, with the largest containing 1732 and the smallest, 2. We used this to see how well small documents were classified by the LDA and LIWC models. To elicit responses containing enjoyable or frustrating content, the interviewer used the following prompt: talk about an enjoyable/frustrating experience that happened in the last week.

Model design

The overall approach to the model design consists of two processing steps: (1) extract the features from each document in the set, and (2) use these features to predict whether the interview described a frustrating (negative) or enjoyable (positive) experience. The difference between our model and LIWC is the feature extractor used in step (1): LDA topic proportions vs. LIWC word count. The design of the LDA feature extractor is shown in Figure 1 and that of LIWC in Figure 2. Prior to extracting features from the document set, the collection is split using a 90/10 proportion into a training and test set,

shown in steps 1 and 2 in both figures. This is to create ”set-aside” test documents, which can be used to evaluate how well the model predicts whether an interview is enjoyable or frustrating for a document that has not been used for training. The training set is used to build the generative topic model which is then used to infer the topic proportions (i.e., features) of both the training set as well as the set-aside test set. Once the training and test sets have been created, feature extraction follows two distinct processes for LDA and LIWC.

LDA training and feature extraction

Once we have split our document collection, we can use the training set to generate the topic model and infer its thematic structure, i.e., topic proportions. We use the Gensim (Řehůřek & Sojka, 2010) implementation of LDA, a robust, stable version that is widely used in academic research for topic modeling and natural language analysis. While it is possible to adjust many of the implementation’s parameters (e.g., the Dirichlet priors for the per document distributions and for the per topic word distributions), we accept the default values. As mentioned earlier when we introduced LDA, the generative model assumes a number of topics over which an initial distribution of documents (i.e., $P(\Theta_d)$) is estimated. We now describe how we selected the number of topics.

Recall that a topic model tries to discover a thematic structure in a document collection; it is trying to find structure in otherwise unstructured text. One of the characteristics of this type of machine learning method is that it does not guarantee that the topics will be interpretable by humans. Thus a measure is needed to automatically evaluate the topic quality of the topics generated by the LDA model. We use the topic coherence pipeline available in Gensim which is an implementation of the method described in (Röder, Both, & Hinneburg, 2015). In the context of topic modelling, a *coherent* model is one in which words are treated as facts; coherence can then be

evaluated on the basis of how well the words in a topic "support" one another, as when we speak of a coherent set of facts. In the topic model, words support one another based on their probability of co-occurring together. The coherence measure produced by the Röder et al. (2015) framework is a real number representing an aggregation of probability estimates; this number can be used to compare the topic quality of different topic models. Röder et al. (2015) report that the model has been extensively compared with human gold-standard coherence measures using Wikipedia as a reference corpus and has performed quite well. Figure 3 shows a plot in which the LDA model was run with an increasing number of topics in steps of 2, from 2 to 100, against which the coherence score was calculated. We can identify eight local maximum values at 4, 16, 24, 34, 44, 50, 64, and 91 topics respectively. We hypothesize that the interview process, during which a participant was asked to recall a frustrating and enjoyable activity, tends to generate a large set of words with different co-occurrences across participant interviews. However, there are a set of topics which distinguish between frustrating and enjoyable content, allowing the model to use these topics to predict emotion valence. We will describe how these eight topic-number values were used in the model evaluation in a subsequent section.

As shown in step 2 of Figure 1, once we have split the interview transcriptions into training and test document sets, we set aside the test set and proceed to pre-process the training set. The purpose of pre-processing is to transform the original text into a more efficient set of words, removing information that does not help the LDA model infer its thematic structure. Pre-processing lemmatizes words (i.e., words in the third person are changed to the first person; verbs are changed to the present tense) and words are stemmed (i.e., reduced to their root). Common "stop" words (e.g., the, is, at) and disfluencies (e.g., um) are removed. Document text is split into sentences and then into

words and word frequencies are computed. It should be noted that during the pre-processing, the ordinal nature of the document structure is broken and it becomes a bag of words. The LDA model does not use grammatical structure to infer thematic structure.

Training is completed once the LDA model has estimated the hidden variables Z (per-word topic assignment) and Θ_d (per-document topic proportion), which the LDA model in Gensim does automatically on our behalf. At this point we have a trained topic model to which we can supply unseen documents and obtain topic proportions; we can also extract the topic proportions already assigned to the documents it used for training. In either case, the topic model produces a set of feature vectors, one for every document the size of each being the number of topics. However, we do not yet have an association between the topic proportions and the classification of a document as "frustrating" or "enjoyable". In a subsequent section, we will describe how we can use a machine-learning tool known as *classifier* to make this association.

LIWC feature extraction

Takahashi and Tickle-Degnen (2010) used five LIWC dictionaries (categories) to measure participants' verbal expression of positive and negative emotion. They are: positive emotion, anxiety or fear, anger, sadness or depression, and achievement. The researchers used the 2010 version of LIWC to extract word counts in these categories from participant interviews. An analysis of variance (ANOVA) was used to show a statistically significant effect that participants used more words categorized by LIWC as expressing positive emotion when talking about enjoyable activities rather than frustrating activities, and used fewer words expressing negative emotion. Alternatively, participants used more negative words when asked to recall a frustrating activity. Thus we chose these five categories to be used as features, hypothesizing that they could be

associated with the classification of an interview as being frustrating or enjoyable. As shown in step 2 of Figure 2, the participant interviews were lightly processed to remove disfluencies and then input to the 2015 version of the LIWC software. The resulting output is a set of feature vectors for every document, each vector of size five and where each feature represents the word proportion of the corresponding LIWC emotion category. The feature vectors generated by LDA and LIWC were then used by a classifier to learn the association between the features and the type of interview.

Using features to predict emotion valence

In machine learning, a classifier is a software tool used to predict classes of items rather than values; the latter is performed using regression techniques. In our model we use a logistic regression (LR) classifier to predict a set of two possible interview classes, *interview = {frustrating, enjoyable}*. We use a stable, widely-used implementation of an LR classifier from Scikit-learn, a free software machine learning library (Pedregosa et al., 2011). Logistic regression, developed by statistician David Cox (1958), computes the probability of output in terms of input and this can be used to construct a classifier by choosing a cut-off probability value (i.e., 50%) and classifying input values greater than the cut-off as one class and below the cut-off as the other. The classifier is trained and used to predict the interview classes in exactly the same way for both the topic features and LIWC features (see Figure 4); the only difference is the feature set used, and the following discussion holds for both sets.

Training the logistic regression classifier consists of finding the parameters θ of the model such that it sets high probabilities for enjoyable content and low probabilities frustrating content. This is achieved by minimizing the cost function, $c(\theta)$, where the probability estimate is $\hat{p}$ and the training label is y:

$$c(\theta) = \begin{cases} -\log(p), & if\ y = 1 \\ -\log(1-p), & if\ y = 0 \end{cases}$$

Consistent with the goal of the classifier, $log(x)$ grows larger when x approaches 0, and therefore, the cost will be large if the classifier estimates a probability close to 0 for an enjoyable interview; likewise, it will also be large if it estimates a probability close to 1 for a frustrating interview. Alternatively, −log(x) is close to 0 when x approaches 1 and the cost will be close to 0 when the estimated probability is close to 0 for frustrating interview and close to 1 for an enjoyable interview. To compute the value of θ that minimizes the cost function, the Scikit LR classifier implementation uses an optimization method known as stochastic gradient descent (a good discussion can be found in Géron, 2017). Once classifier training was completed, we evaluated the LDA and LIWC models' performance using materials from (Tickle-Degnen et al., 2010).

Results and evaluation

Experiment 1: Predicting interview class using larger word counts

For this evaluation, we used the document collection, from the *Self-management Rehabilitation and Health-Related Quality of Life in Parkinson's disease* database where the average word count = 303 to train and test the model; there are 332 and 34 documents in the training and test sets respectively. The LDA feature extractor (see Figure 1) was trained eight times by setting the LDA model's parameter for the number of topics according to the eight values identified by the coherence model as local maxima (see Figure 3). Each training session i, where $1 \leq i \leq 8$, and $n_i = \{4,16,24,34,44,50,64,91\}$ topics generates a feature vector of size n_i for each document in the training set. Each feature vector is associated with a document's target

label (i.e., $enjoyable = 1, frustrating = 0$) and the $(feature, target)$ pair is used to train the logistic regression (LR) classifier using a method known as *K-fold cross validation*. The results for each training session are shown in Table 1. In K-fold cross validation, the training set is split into K distinct subsets called *folds*. We set K = 10; this is typical for the size of our training set, which is considered small compared to typical machine learning problems that can have several thousand training instances. This process trains and evaluates the LR classifier ten times choosing a different fold for testing every time and training on the remaining nine folds.

We include more robust metrics than accuracy to evaluate the model performance: precision, recall, and F_1. Precision gives a measure of the accuracy of positive predictions. It is computed as shown in the equation below, where *TP* is the number of true positives and *FP* is the number of false positives. Thus, a model with low precision will tend to signal a high number of "false alarms". It is often compared with another measure called *recall*, also known as sensitivity or the *true positive rate*, i.e., the proportion of positive instances correctly identified by the model. It is computed as shown below, where *FN* is the number of false negative instances.

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN}$$

For example, referring to Figure 1, when using four features and when the model predicts the interview is enjoyable, it is correct only 56% of the time; when an interview is actually enjoyable, it predicts so 56% of the time. It is common when evaluating classifiers to combine precision and recall into a single statistic, called the F_1 score. This score is the harmonic mean of the precision and recall, which unlike the arithmetic mean, balances both; you cannot get a good F_1 if either are low. F_1 gives its best score at 1, when precision and recall are perfect. Thus, the harmonic mean will generate high F_1

values when both the precision and recall are high. We can see, for example, that the F_1 score of 61 is highest when features = 24, 34.

As discussed, we have trained eight LDA feature extractors, corresponding to the number of topics we presented as a parameter to the model and that we have set-aside 10% of our documents (i.e., 34) which have never been used to train either the LDA feature extractor or the LR classifier. For each of these feature extractors, we present the test documents in order to extract their features and then we present them to our trained classifier which predicts whether the documents are either frustrating or enjoyable interviews (refer to Figure 4). The results are shown in Table 2 which gives the classification accuracy for each feature set size used. The accuracy is the mean score across the 34 test documents.

The five emotion categories used by Takahashi and Tickle-Degnen (2010) were used to extract features from the 332 training documents as shown in Figure 2. These feature vectors were paired with their corresponding document target labels and we followed the same 10-fold cross validation procedure described in the previous section to train the LR classifier (refer to Figure 4). The precision, recall, and F_1 evaluation metrics are shown in Table 1. We then used the LIWC categories to extract the features from the set-aside test documents and presented them to the trained LR classifier in order to predict each document's interview category. The results are shown in Table 2 which gives the mean classification accuracy across the 34 test documents.

Experiment 2: Predicting interview class using smaller word counts

In this experiment, we investigated how well an LDA model trained on a collection of 404 documents from the same database whose word count ranged from 2 to 1732, with an average of size of 258 could accurately predict the interview class using test

documents representing short bursts of dialog. For the test set, we used transcripts that were edited to obtain the first 20 seconds of conversation. These documents have an average word count of 13 words, and ranging from 22 to 2 words; this is the typical word count found in the average sentence. The LDA feature extractor was trained using number of topicsequal to 4.

As can be seen in Table 3, the LDA model F1 score of 0.80 using the test set is considerably better than the LIWC model score of 0.44. Presumably this is because it has extracted the topics from the context of all the documents in the training set and thus is able to use this information to situate the unseen document in these topics. In contrast, LIWC only uses the words in the test document presented to it, which may contain insufficient content to accurately predict the emotion content. We do not report statistics for the training process since the training data is similar to what was used for Experiment 1 and the purpose was to evaluate short test documents.

Discussion

The findings suggest that LDA can be used to discover a set of topics whose proportions for any given document in the collection can be used to parsimoniously represent the positive or negative emotion content of that document. It appears that the number of topics used to extract the features does not greatly affect the performance of the classifier. The most compact feature set using four topics, had an F_1 score of 0.58 and resulted in a test set accuracy of 71% whereas the largest feature set of 91 topics had an F_1 score of 0.61 and a test set accuracy of 74%. In comparison, the LIWC model used five categories of words which have been previously shown to correlate with interview category (Takahashi & Tickle-Degnen, 2010). Words belonging to each category were tallied and used to calculate their category's proportions. This approach produced an

F_1 score of 0.74 and a test set accuracy of 76%. When faced with a number of choices of increasing complexity, all of which have similar explanatory power in a model, it is reasonable to choose the least complex explanation (i.e., Occam's razor). The least number of topics that can be used to train the LDA model and generate features that separate the documents into positive and negative emotion categories is four and in the remainder of the discussion, we will assume this version of the LDA feature extractor.

Classification accuracy is only one part of the evaluation; precision and recall, described earlier, are metrics that provide more nuanced on the models' predictive behaviour. In experiment 1, precision and recall are both 0.56 during the cross-validation training. This gives us insight as to how the model will perform across a variety of test sets. These metrics suggest that LDA does better than chance predicting the interview category in two situations: (1) when it makes a prediction, it is correct 56% of the time (precision) and (2) given the actual target, the model makes the same prediction 56% as well (recall). The LIWC model, with an F_1 score of 0.74, performs better in both cases, 74% of the time. We note that for both models, precision equals recall and thus neither is biased toward giving more false positives (precision) or false negatives (recall). These statistics suggest that the LIWC model would more accurately classify new interviews whose average word count is 303 words (approximately 22 sentences per document on average). The performance differential may be due to the five dictionaries selected to be the source of the features used in the LIWC model.

On the other hand, in situations where the model is likely to encounter one or two sentences, Experiment 2 (Figure 3), which produced an F_1 of 0.80, suggests the LDA to be the better choice. We might, for example, create a communication-assistive tool to infer the emotion content of each interaction between a person with PD and a caregiver, which most likely consists of short bursts of dialog. In this experiment, the

precision and recall metrics have different values in both LDA and LIWC. In LDA precision = 0.75 and recall = 0.86 suggesting that the model is biased away slightly from making false negative predictions and more towards false positives. In an interaction with a PD person and a caregiver, the LDA model will evaluate that turn in the conversation to contain more positive emotion (more enjoyable) when it may in fact contain more negative emotion (more frustrating), slightly more frequently (11%) than making a negative prediction when the turn is positive. However, the LIWC precision of 100% suggest that it will almost never make a false positive prediction, but when it does makes an incorrect prediction, which it did in this experiment, 1.00 - 0.64 or 36% of the time, it is likely to be a false negative. A higher level of false negative predictions (predicting frustrating content when the actual is enjoyable) is an example of the desynchronization of mental state and facial expression which Tickle-Degnen, Hall, and Rosenthal (1994) report "can have a profound effect on communication ability and quality of life".

Another characteristic of the LDA model is that it is not sensitive to the language of the text and uses the entire document collection during training to infer the hidden topic structure. Thus it can learn thematic structure from documents in any language and use what it has learned to place documents it has not seen in the topic structure, even short, one or two sentence documents. LIWC, however, would have to be modified to incorporate a new language and its dictionaries would have to be updated to ensure the new language's words were placed in the proper emotion categories. Thus we hypothesize superior performance of LDA in spoken dialog of persons with PD. Persons with PD have difficulty with enunciation and voice volume, therefore automated speech recognition technology (ASR) at its current level (i.e., 15% *Word Error Rate),* is likely to produce inaccurate transcripts. This will make it difficult for

LIWC to recognize words and process word counts. Since LDA is not sensitive to the word orthography, it should still be able to use the imperfect transcriptions to extract the topic features; this theory however remains to be tested. At present, we used the model to classify emotion valence categorically as enjoyable (positive) or frustrating (negative); however it is also possible to use the model to predict other discrete emotions including the level of arousal. In a future version of the model, predicting both valence and arousal can be used to observe and inform the emotion trajectory of a conversation as it unfolds between any two participants, for example, in therapy or counselling sessions.

Limitations

The results of this study suggest that topic modeling could extract features associated with emotion valence using verbal transcriptions of interviews in which participants were specifically asked to recall an enjoyable and a frustrating experience. We shall point out a few limitations of this approach. People living with PD might have talked about frustrating things when describing enjoyable experiences, this sometimes happens in the context of chronic illness, especially with people with depression. However, our participants were screened for depression and were found to not be clinically depressed.

Also, human emotional state can change in far more complex ways and in more subtle gradations than the positive/negative emotional categories we have explored in this study. In addition, further research in how humans combine input from several modalities such as visual, auditory, and tactile to generate understanding of complex mental states such as embarassment, thinking or depression is needed to inform a more naturlistic and incremental model. Furthermore, the topic model infers the hidden thematic structure and the topics are not easily interpretable. The topics cannot really tell us much in a way that makes sense to a human why a certain text has an enjoyable

or frustrating emotional content. Finally, our training and testing dataset was comparatively small. At most, we had 448 transcripts in the document collection, a limited amount of data compared to typical machine learning endeavours which may have thousands of training examples available. This means that in order to generalize the test results to a domain beyond that described in Tickle-Degnen, et al. (2010), additional training documents will have to be used.

The initial purpose of this research was to investigate whether interview data from persons with PD could be used to train a model to predict whether a new utterance described something enjoyable or frustrating. We view this as a first stage in building an assistive tool which a caregiver could use to infer the emotional state of a person with PD. In order for the model to be useful in a clinical setting as a communication assistive tool, it will have to be developed further to generate the more incremental and subtle gradations of human emotional states. In addition, clinical trials would need to be conducted to assess its usefulness.

Conclusion

In this article, we investigated an automated method for inferring the emotional state of a person with Parkinson's disease using a machine learning approach. Our results show that the LDA model performs better with shorter text which makes it more suitable for evaluating emotional content of short dialog turns. For longer documents, LIWC performs better; however, it has the shortcoming of assuming a constant, non-evolving language and is dependent on manually selecting the dictionaries to be used as features in the problem domain. It is also non-generalizable to other languages for which dictionaries have not yet been created.

The LDA model is a first step towards creating an assistive tool which the caregiver can use to infer the emotional state of a person living with PD. Given that

many people with PD live in the community, the caregiver is likely to be a member of the family who is not necessarily trained or accustomed to the symptoms of the disease and may make incorrect inferences about the person's true emotional state. Thus an assistive tool equipped with the ability to accurately and immediately provide feedback on the emotion content of a conversation is not only beneficial for improving the social interaction with the PD patient, it can improve the quality of life in the home, including that of the family members. This capability can also be to assist the rehabilitation therapist in, for example, during client evaluation, helping preserve the client's dignity in situations where the client's claims to be happy is belied by her affectless face. This technology is not restricted to the domain of people living with Parkinson's disease; it should be able to generalize and serve as an intelligent agent useful for monitoring the emotional content of the interaction between any two parties, providing real-time feedback on the emotional content as the interaction unfolds.

References

Blei, D. M., Eng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Language Research*, 3, 993–1022.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society*. Series B (Methodological), 20(2), 215–242.

DeGroat, E., Lyons, K. D., & Tickle-Degnen, L. (2006). Verbal content during favorite activity interview as a window into the identity of people with Parkinson's disease. *Occupational Therapy Journal of Research: Occupation, Participation, and Health*, 26(2).

Géron, A. (2017). *Hands-on Machine Learning with Scikit-Learn & TensorFlow* (1st ed.). Sebastopol, CA: O'Reilly.

Goetz, CG, Poewe W., Rascal, O., et al. (2004) Movement Disorder Society Task Force report on rating scales for Parkinson's disease. *Movement Disorder*, 19(9), 1020-1028.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Web and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. Damerau (Eds.), *Handbook of natural language processing* (2nd ed.). San Rafael, CA: Chapman and Hall.

Ma, H., Saint-Hilaire, M., Thomas, C. A., & Tickle-Degnen, L. (2016). Stigma as a key determinant of health-related quality of life in Parkinson's disease. *Quality of Life Research*, 25(2), 3037–3045.

National Institute of Health (2018). *Parkinson's disease: Challenges, progress, and promise*.. https://www.ninds.nih.gov/ Disorders/All-Disorders/Parkinsons-Disease-Challenges-Progress-and-Promise. (Accessed: 2018-10-29)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. doi: 10.15781/T29G6Z

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modeling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). New York, NY, USA: ACM. http://doi.acm.org/10.1145/2684822.2685324.

Sturkenboom, I. H., Graff, M. J., Borm, G. F., Adang, E. M., Nijhuis-van der Sanden, M. W., Bloem, B. R., & Munneke, M. (2013). Effectiveness of occupational therapy in Parkinson's disease: study protocol for a randomized controlled trial. *Trials*, 14(34). doi: 10.1186/1745-6215-14-34

Takahashi, K., & Tickle-Degnen, C. W. J. L. N., L. (2010). Expressive behavior in Parkinson's disease as a function of interview context. *American Journal of Occupational Therapy*, 64(3), 484–495.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tickle-Degnen, L., Ellis, T., Saint-Hilaire, M., Thomas, C., & Wagenaar, R. C. (2010). Self- management rehabilitation and health-related quality of life in Parkinson's disease: A randomized controlled trial. *Movement Disorders*, 25, 194–204.

Tickle-Degnen, L., Hall, J., & Rosenthal, R. (1994). Nonverbal behavior. *Encyclopedia of Human Behavior*, 3, 293–302.

World Health Organization (2006). *Neurological disorders: Public health challenges*. https://www.who.int/mental_health/neurology/neurodiso/en/. (Accessed: 2019-03-28)

Table 1. Experiment 1: Model evaluation using 10-fold cross-validation for 332 training documents with average word count = 303; max = 1732; min = 131

LDA evaluation

Features	Precision	Recall	F_1
4	0.56	0.56	0.58
16	0.54	0.55	0.54
24	0.62	0.62	0.62
34	0.62	0.63	0.62
44	0.59	0.59	0.59
50	0.57	0.57	0.57
64	0.57	0.58	0.57
91	0.63	0.63	0.61

LIWC evaluation

Features	Precision	Recall	F_1
5	0.74	0.74	0.74

Table 2. Experiment 1: Model testing using with average word count = 303; max = 1732; min = 131

LDA evaluation

Features	Accuracy
4	0.71
16	0.65
24	0.59
34	0.68
44	0.65
50	0.53
64	0.56
91	0.74

LIWC evaluation

Features	Accuracy
5	0.76

Table 3. Experiment 2: Model testing using 14 documents with average word count = 13; max = 22; min = 2

LDA evaluation

Features	Accuracy	Precision	Recall	F_1
4	0.79	0.75	0.86	0.80

LIWC evaluation

Features	Accuracy	Precision	Recall	F_1
5	0.64	1.00	0.29	0.44

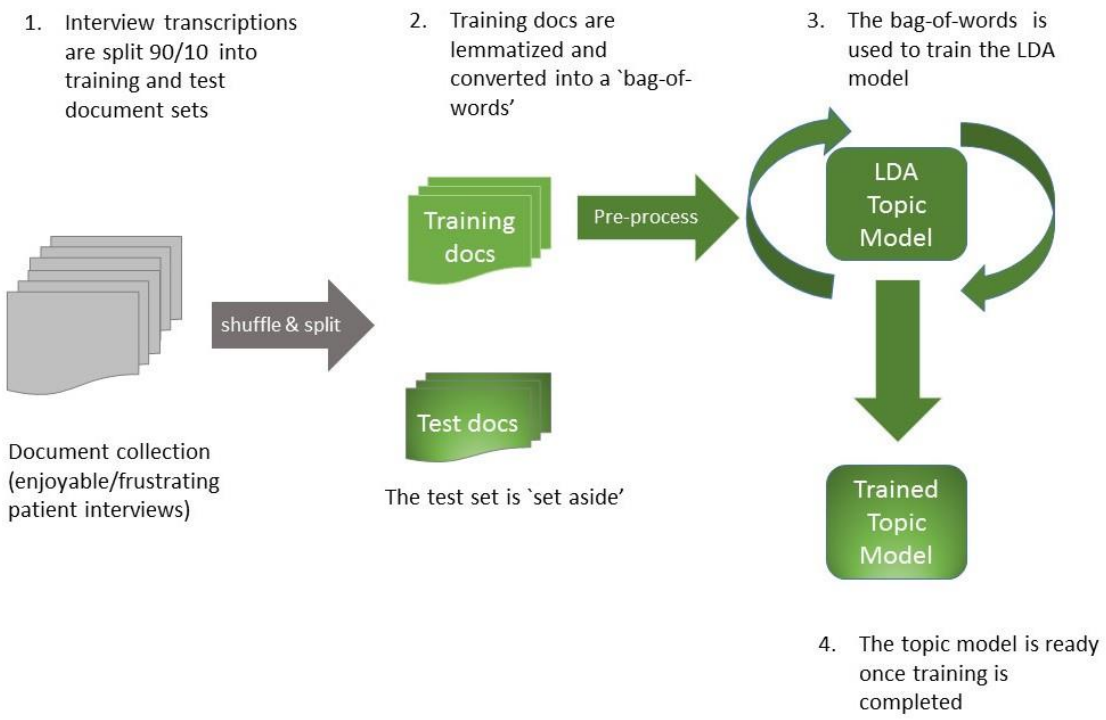

Figure 1. LDA feature extractor.

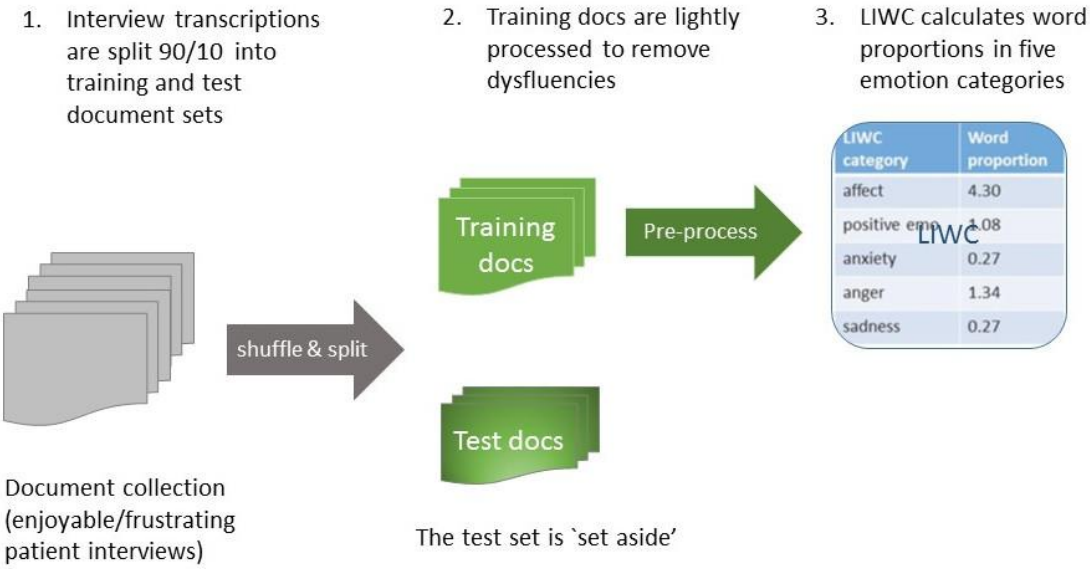

Figure 2. LIWC feature extractor.

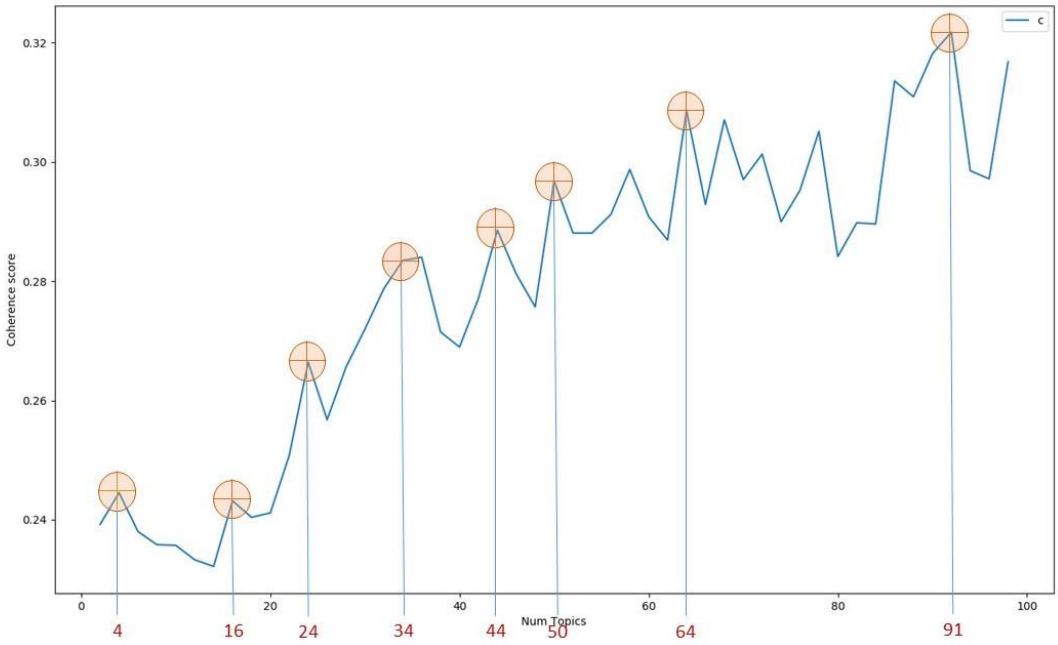

Figure 3. Coherence score by number of topics.

The LDA model was trained repeatedly using the training set, starting with 2 topics. At the end of each iteraction, the coherence score was calculated and the number of topics was increased by 2 until 100 topics was reached. Eight local maxima are identified by cross-hairs on the graph.

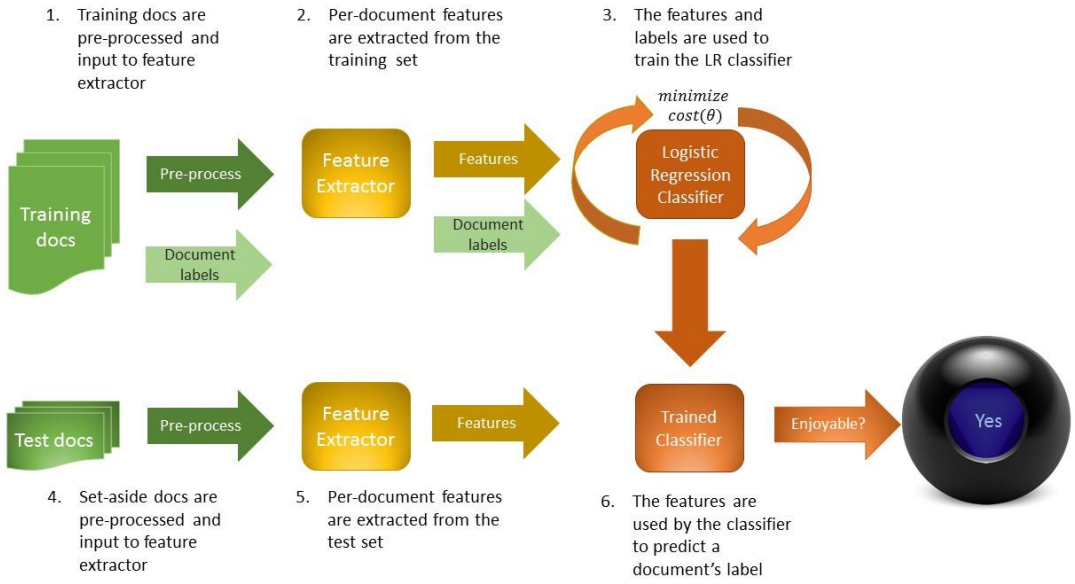

Figure 4. Logistic regression classifier.

(1,2) Features are extracted from the training set using either LDA or LIWC. (3) Training proceeds until the model minimizes a cost function, $c(\theta)$, which penalizes misclassification. (4,5,6) Features are extracted from the set-aside documents and presented to the trained classifier for predicting the interview type (yes = enjoyable, no = frustrating).