Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Modeling Latent Dimensions of Human Beliefs

Huy Vu, [1] **Salvatore Giorgi,** [2] **Jeremy D. W. Clifton,** [3]
Niranjan Balasubramanian, [1] **H. Andrew Schwartz** [1]

[1]Computer Science Department, Stony Brook University,
[2]Department of Computer and Information Science, University of Pennsylvania,
[3]Positive Psychology Center, University of Pennsylvania
{hvu, niranjan, has}@cs.stonybrook.edu, {sgiorgi, cliftonj}@sas.upenn.edu

Abstract

How we perceive our surrounding world impacts how we live in and react to it. In this study, we propose LaBel (Latent Beliefs Model), an alternative to topic modeling that uncovers latent semantic dimensions from transformer-based embeddings and enables their representation as generated phrases rather than word lists. We use LaBel to explore the major beliefs that humans have about the world and other prevalent domains, such as education or parenting. Although human beliefs have been explored in previous works, our proposed model helps automate the exploring process to rely less on human experts, saving time and manual efforts, especially when working with large corpus data. Our approach to LaBel uses a novel modification of autoregressive transformers to effectively generate texts conditioning on a vector input format. Differently from topic modeling methods, our generated texts (e.g. "the world is truly in your favor") are discourse segments rather than word lists, which helps convey semantics in a more natural manner with full context. We evaluate LaBel dimensions using both an intrusion task as well as a classification task of identifying categories of major beliefs in tweets finding greater accuracies than popular topic modeling approaches.

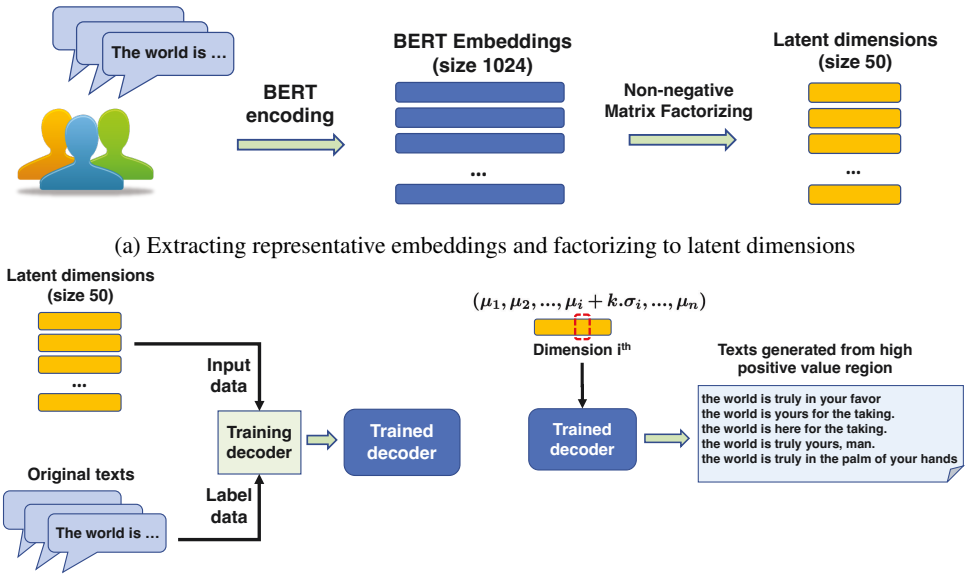
Introduction

Our perceptions of the surrounding world have a great impact on how we live and respond to it. For example, one who thinks "the world is an open book, full of opportunities" might be very interested in learning about and traveling to new places, or taking on new opportunities. On the other hand, one who thinks "the world is always against me" might think that most things in life are discouragingly difficult and avoid challenging situations. Indeed, human perceptions or general beliefs have been a long-running research topic for psychology and social science, with recent work turning to Twitter, among other sources, to better understand general categories of "primal world beliefs", or "primals". (Clifton et al. 2019; Clifton 2020; Stahlmann et al. 2020). In these studies, primal world beliefs were explored by having experts read through a large collection of texts comprising of sacred texts, novels, speeches, treaties, films text and thoroughly analyze human's major beliefs about the world. Such studies, however, have relied mostly on manual efforts from experts, hence largely restricted in scale. This suggests the need for a method that is more automatic and requires less human effort.

In this paper, we seek to automate the process of deriving dimensions of general beliefs about the world, and doing so in a way where such dimensions are easily interpretable and thus expand the scope of social belief tracking to consider more perspectives of more diverse people. We achieve this by using machine learning models to analyze large corpora of text to explore latent major beliefs statistically. Although this problem is quite in line with topic modeling problems solved with methods such as Latent Dirichlet allocation (LDA), we suspect that the way topic modeling represents each latent dimension as a set of words might limit their ability to fully capture or express the semantics of latent beliefs due to the discrete, context-lacking nature of words. Therefore, we suggest an alternative method to topic modeling to explore and represent latent semantics from texts, by generating natural texts with full grammar and contexts, that is potentially more interpretable than discrete words resulting from topic modeling methods.

We propose *LaBel* (Latent Beliefs Model), which both captures latent dimensions of beliefs and, importantly, provides the ability to express such latent dimensions as generated natural language rather than word probabilities. LaBel takes advantage of strong pre-trained contextualized BERT embeddings to rely less on word co-occurrence or uncontextualized word embeddings, and then uses a novel accompanying decoder to generate sample descriptions of the latent dimensions (e.g. "The world is a safe place"). We compare LaBel to a variety of topic modeling techniques in terms of interpretability and utility for classifying beliefs, as well as show the generalizability of the method over a variety of beliefs such as beliefs about education or parenting.

The contributions in this study include: (1) the exploration of major beliefs expressed by people about the surrounding world using an automatic computational pipeline, (2) a novel alternative to topic modeling allowing latent semantic dimensions to be represented by generated phrases rather than lists of words, (3) the demonstration that latent dimensions represented as generated phrases can be more interpretable than representing dimensions with their most prevalent words as is common in traditional topic modeling. Ad-



(a) Extracting representative embeddings and factorizing to latent dimensions

(b) Training decoder generating texts from factorized embeddings

(c) Interpreting latent dimensions by generating texts from high positive values range

Figure 1: Overview of LaBel, the proposed method to explore people's major beliefs from social media texts.

ditionally, we also (4) explore different approaches for modifying autoregressive transformer models to generate texts conditioned on a vector rather than leading tokens or with prompting texts as in other related works (Pilault et al. 2020; Wolf et al. 2019; Zhang et al. 2020; Adelani et al. 2019).

Related Work

Previous work on human beliefs about the world has used manual coding techniques to identify 26 primal world beliefs grouped into 3 levels: primary level (including 1 primal *The world is Good*), secondary level (3 primals *The world is Safe*, *The world is Enticing* and *The world is Alive*), and tertiary level (22 primals including *The world is Harmless*, *The world is Meaningful*, *The world is Stable*, etc.). A primal world belief is defined to strictly have the following six criteria: Simple, Adjectival, Goal-relevant, Maximally General, Automatic, and Active (Clifton et al. 2019). Our work was inspired by these studies but instead seeks to automatically identify latent dimensions of major beliefs, applied to any targeted topic or point of view (e.g., the world, politics, parenting, and education).

"Belief" is a psychological concept involving cognitive construct. However, as many other psychological traits such as personalities, one of the most effective ways to perceive and evaluate them is through language usage, such as written texts. This is also the approach used in the original work exploring primal world beliefs (Clifton et al. 2019), in which beliefs are found by analyzing texts collected from a variety of sources, including sacred texts, novels, speeches, social media texts, that cover many aspects of human lives along history. Our work is premised on this idea to analyze major beliefs from a large corpus of social media texts, which in our opinion, reflect more contemporary and more individual beliefs about the world than the corpora used in previous works. Our automatic method can still easily be applied to any text corpus of any size, while saving time and manual efforts costs.

Considering the goal of finding interpretable latent dimensions from texts, LaBel is similar to topic modeling. However, our method contributes two technical novelties compared to traditional topic modeling methods. First, we take advantage of strong pre-trained contextualized embeddings to analyze the texts which provide richer information than word co-occurrence or uncontextualized word embeddings, such as those used in LDA (Blei, Ng, and Jordan 2003), biterm (Yan et al. 2013), pLSA (Hofmann 1999), or ETM (Dieng, Ruiz, and Blei 2020). This rich information is particularly helpful for modeling short text datasets such as tweets. Second, traditional topic modeling methods explore the topics and represent them with a collection of representative words. Although this representation offers a holistic view of the topic, it might be difficult for readers to assemble the words and figure out the semantics of the topic due to the lack of grammar and context, as in a complete sentence that we as humans are used to reading and understanding. Our study suggests that instead of describing the explored topics with discrete keywords, we can build a decoder to directly generate texts describing these topics. This representation is potentially more understandable since humans are familiar with reading and absorbing information from complete, full-context sentences.

One might also suggest that in order to describe topics with complete sentences rather than words, we can simply sample tweets having high scores in each latent dimension. However, our generative method has the advantage of (1) better generalizing the topics' semantics through learning

and (2) the ability to generate original text at any point in the continuous latent space. Additionally, sampling the dataset for representative tweets raises privacy issues when analyzing a user's exact tweets.

Regarding the decoder used in our model, which will be described in more detail below, our proposed method uses the modified GPT-2 (Radford et al. 2019) model to generate text conditioned on factorized embedding vectors as shown in Figure 1c. Although pre-trained autoregressive transformer models like GPT-2 have been used as a conditional texts generative model for many applications such as abstractive summary (Pilault et al. 2020), dialog generation (Wolf et al. 2019; Zhang et al. 2020), sentiment-preserved products review generations (Adelani et al. 2019), and emotion grounded text generation (Santhanam and Shaikh 2019), most of these studies have the conditions represented as prompting text or special leading signal words or phrases. In this work, however, we want to fine-tune GPT-2 to generate texts from the continuous vector format, which, to the best of our knowledge, has not been studied. Therefore, we develop a modification of GPT-2 to achieve this goal, which will be described in the following section. The method in Bowman et al. (2016) also generates texts from continuous embeddings, although it requires training a model from scratch rather than fine-tuning that takes advantage of well pre-trained models such as GPT-2.

Notice that although we built our model upon the GPT-2 model in this work, our proposed modification method can be used for almost all current state-of-the-art text generative model, including GPT-3 (Brown et al. 2020), XLNet (Yang et al. 2019), and BART (Lewis et al. 2020). These models are also built upon the transformers (Vaswani et al. 2017) building block, and hence, can be modified straightforwardly with our proposed approach for GPT-2. The main reason we used GPT-2 in this paper is that it is more common and widely known than its counterparts as one of the first effective text generative models. Besides that, it is relatively smaller, having fewer parameters, hence requiring less training cost, which allows our method to be more easily reproduced.

Latent Belief Model

Our goal is to design a generative model that enables the exploration of latent beliefs. In particular, we want to organize different types of beliefs along latent dimensions in an embedding space and use a generative decoder that allows for flexible exploration of this latent space by producing textual representations of various points in this latent space.

In designing our latent belief model LaBel, we want to exploit the powerful encoding and generation capabilities of large pre-trained language models as decoders, but also devise mechanisms that will allow these decoders to generate text from any given point in an external latent space. Figure 1 presents the overview of our implementation of such a latent belief model. First, we collect tweets describing people's thoughts about the world by matching tweets against simple look-up expressions (e.g. using the prefix "the world is ...") that have been used in previous research on primal world beliefs (Clifton et al. 2019). We then extract BERT embeddings from these tweets to use as their semantic representations. Then, we use Non-negative Matrix Factorization (NMF) to project these representations down to a low-dimensional latent space. Finally, we use a GPT-2 based decoder to generate texts as a way to interpret these latent dimensions. We introduce a modification to the GPT-2 model to allow it to decode from any point in the latent belief space and train it using the reconstruction task.

Latent Belief Embeddings

We obtain the latent belief embeddings from contextualized representations of the collected tweets. First, we process the collected tweets through the BERT-Large model (Devlin et al. 2019). For each tweet, the contextualized embeddings of each word are averaged across the last four layers, and then averaged across the sentence to have one embedding vector of size 1024. This practice of integrating embeddings from the last layers of BERT model as representations has been shown to be useful for many downstream tasks for social media texts (Matero et al. 2019; Vu et al. 2020; Ganesan et al. 2021). Then, we project these vectors down to a low-dimensional belief embedding space using the non-negative matrix factorization algorithm.

Non-negative matrix factorization (Lee and Seung 2000) produces factors with only positive latent dimension values. This makes its dimensions easier to interpret than those from other standard dimensionality reduction methods and, hence, is a common choice for topic modeling with bag-of-words features. In this work, we use NMF to factorize and project the BERT embeddings from 1024 dimensions down to 50 dimensions, which roughly corresponds to the number of primal world beliefs studied in Clifton et al. (2019). Formally, the BERT embeddings matrix is first preprocessed by subtracting their minimum value from them to create an all non-negative matrix X (size $[n_{tweets} \times 1024]$). The NMF algorithm is then used to produce two non-negative matrices W and H, which minimize the following reconstruction error:

$$||X - W.H||_F \quad (1)$$

with $||.||_F$ is the Frobenius norm.

Here H is the coefficients matrix (size $[50 \times 1024]$) and W (size $[n_{tweets} \times 50]$) is the resulting *latent belief embedding*, the low-dimensional projection of the original BERT embeddings. Each dimension in the latent belief embedding is expected to correspond to one specific type of belief about the world. The number of dimensions is set to 50 in order to closely match the total number of primals explored in (Clifton et al. 2019) — 26 categories, each with two poles (e.g. Beautiful world belief has two poles of "beautiful" vs "ugly"). Note that although we used the 26 categories with two poles to justify the choice of 50 dimensions, the beliefs explored by our method don't necessarily need to match all the ones found in (Clifton et al. 2019). This is because we work on a slightly different dataset — social media texts, which as mentioned above, might contain more contemporary and individual beliefs. The NMF algorithm is implemented with the *scikit-learn* Python library, using the following hyperparameters: tol = 1e$-$4, max_iter = 200, $init$

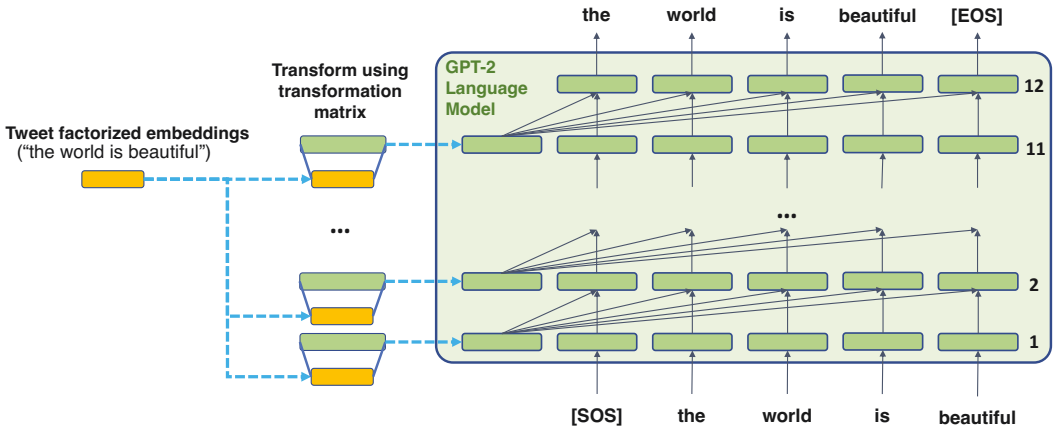


Figure 2: Decoder architecture used to generate texts conditioned on factorized embeddings vectors. Yellow box represents input latent embedding vector. Green box represents hidden state representation vector.

= "$random$", $n\_components$ = 5 (other hyperparameters are kept as defaults). In the Supporting Information files, we also tested our model with different values of number of dimensions, such as 25 or 100, to support the robustness of our approach.

Interpreting Latent Dimensions with Generated Texts

We introduce a new method to interpret the dimensions of the latent belief embeddings by directly generating sentences from these latent dimensions. In prior work on topic modeling, it is typical to represent and explore the latent topics using a collection of keywords that are highly correlated with each topic. While this provides a holistic view of the topic, it can be difficult to understand some keywords without seeing them in a context (e.g. in a sentence) and only gives a single view of the entire topic. However, if we had the capability to represent the latent dimensions through sentences and had a way to generate these sentences for any given point in the latent embedding space, then we can provide a much richer, contextual exploration of the embedding space. To this end, we leverage an existing generative model, GPT-2, and modify it to generate text over the latent embedding space.

Latent Embeddings Decoder. Our goal is to modify GPT-2 to generate text conditioning on the latent belief embeddings.The main challenge here is that the language models like GPT-2 are auto-regressive models, which can be trained to function as a conditional language model but in most cases generation is conditioned on prompt texts or special beginning words or phrases (Adelani et al. 2019; Pilault et al. 2020; Santhanam and Shaikh 2019; Wolf et al. 2019; Zhang et al. 2020). They can't directly be used to condition on vector embeddings in some arbitrary space, as is the case with our latent embeddings. To address this, we introduce a novel latent embedding decoder which includes a simple and effective modification of the standard GPT-2 language model to condition over a vector of continuous variables.

To inject the latent embedding vector as input information into the GPT-2 model, we must somehow find a way to link it to the hidden state representation of the model. Usually, the representation of the first token can be used to capture this input information. However, due to the discrepancy in vectors' size - which is 50 for our latent embedding vector, and 768 for hidden state representation vector, this practice is not trivial to achieve. Hence, we proposed using a linear transformation - a trainable weight matrix - to project the latent embedding vector onto the larger hidden state vector required by GPT-2. To be more particular, at each layer i^{th}, we add to the model a transformation matrix $W^i_{transformation}$ (size 768×50) to transform the embedding vector $d_{latent\_embedding}$ (size 50) to the hidden state vector $d_{transformed}$ (size 768):

$$d_{transformed} = W^i_{transformation} \times d_{latent\_embedding}$$

We feed the embedding vector to all 11 layers of GPT-2, except the last layer since the first token at this layer does not affect the texts generated. The model learns a different transformation matrix for each layer since the information at each layer can be different. Figure 2 illustrates this latent embedding decoder.

We train this decoder using the reconstruction task, where we first obtain the latent embeddings of an original tweet from our dataset and then ask the decoder to regenerate the original tweet from the latent embedding. This training mechanism is therefore similar to how the original GPT-2 is trained, with only one difference being that the sample input will have the additional latent embeddings vector. Figure 2 illustrates the training process of our model for one sample. Once trained, we can now use this decoder to generate tweets for any latent embeddings values – for example sampling from the extreme values of each dimension.

In this paper, the version of GPT-2 used is the base model, containing 12 transformers layers, 12 attention heads, 768 hidden dimensions. In total, our proposed model has 125 million trainable parameters. More thorough details about GPT-2 base configuration can be found in

the original paper (Radford et al. 2019). We trained the model on 16,752 tweets, with a batch size of 48, using learning rate $5e{-}5$ (manually searched and chosen from $[5e{-}3, 5e{-}4, 5e{-}5, 5e{-}6]$). We trained for 10 epochs, using 3 NVIDIA TITAN Xp GPUs.

Generating Texts to Explore Latent Dimensions. The trained decoder provides a way to generate text (belief tweets) that corresponds to any point in the latent embedding space. This provides us with a way to systematically explore the latent beliefs. For each dimension, we collect the distribution of values from the training data, then pick certain points along this distribution and ask the decoder to produce texts corresponding to these values.

To explore any particular dimension, we first fix the other dimensions' values to their mean values. Then, we select a range of values that lie between twice and four-times the standard deviation from the mean, in order to account for the most variation in the dimension without including too many outliers. For example, if we are exploring dimension i^{th}, we will take values at $\mu_i + k \times \sigma_i$ with k ranging from $k = 2.0$ to $k = 4.0$, while keeping other dimensions at mean values, resulting in a point in the embedding space that is shown below:

$$(\mu_1, \mu_2, ..., \mu_{i-1}, \mu_i + k \times \sigma_i, \mu_{i+1}, ..., \mu_n) \quad (2)$$

We then use the aforementioned trained decoder to generate texts directly from these vectors corresponding to each dimension of interest. Because we restrict the values of the other dimensions to their mean values, we make sure the decoder generating texts is conditioning mainly on the interested dimensions.

Evaluation

With the goal of building a model to automate exploring latent dimensions of general beliefs by learning and generating natural texts explaining them, there is no single evaluation to fully determine its performance and usefulness. Therefore, as an early work to attempt this, we lay a foundation for a multi-modal evaluation of the generative results through (a) qualitatively examining the generated texts (b) testing cluster consistency with human evaluation, (c) predicting existing primals with embeddings, and (d) generalizing capability to other topics ("education", "parents"). The following subsections describe the purpose and details for each evaluation method.

Dataset

We focus on the domain of social media, where people regularly share beliefs about the world. Particularly, we collect tweets from Twitter where users can freely express their thoughts and opinions as well as events happening in daily life. An individual's mindset, personality, or even mental health state can also be learned from their social media content (Schwartz et al. 2013; Coppersmith et al. 2015; Kulkarni et al. 2018; Matero et al. 2019). In this paper, since we want to explore people's major beliefs collectively from many individuals, social media texts such as Twitter posts are a promising source of information.

We collected a random 1% sample of Twitter posts from October 2011 and December 2013 (2.24 billion tweets in total) using the official Twitter Streaming API. We filtered out tweets that are not English. As our focus is not on finding particular tweets, we use a simple but relatively precise pattern to identify beliefs about the surrounding world, we filtered to tweets beginning with the phrase "the world is" (e.g., "the world is yours if you want it to be", "the world is against me!", "the world is a fantastic place"). The total number of tweets collected with this filter was 16,752. From each matching tweet, we extract the direct consequence of "the world is..." in the matching sentence. Other sentences in the tweets are omitted to avoid noisy, non-belief context.

In order to focus on original tweets written by users, we remove quoted tweets from other users or famous people (e.g. the quote "the world is a book, and those who do not travel only read a page." by Saint Augustine). We detect these tweets by measuring their length and their frequencies. If a medium to long tweet (defined as exceeding 10 words) is repeated more than 15 times, it is very likely that these tweets are quoted from one original source. For these cases, we reduce their duplicates by taking only a fraction (e.g. 20%) of its total repetition in the dataset. This is meant to enable the belief to still influence the results (i.e., its repeatability is not without merit) but not over-rule original content.

The codes and data for recreating our models are attached in Supporting Information files. Upon acceptance, they will be made publicly available on Github.

Qualitative Examination

Table 1 shows interesting examples of texts generated from different dimensions. We observe that the texts generated for each dimension correspond to a specific point of view about the world. For instance, dimension 19^{th} seems to correspond to the view that life in this world is cruel, merciless, and unbearable. Dimension 28^{th} seems to correspond to the view that the world is always changing, not static. Dimension 29^{th} seems to correspond to the view that this world is very unstable, crazy, or unpredictable. Dimension 32^{nd} seems to correspond to the view that this world is yours, you can control your life, or there are many opportunities available out there for you to try. Dimension 46^{th} is an interesting case that corresponds to the view that this life is not real but more like a simulation that exists inside our heads.

Notice that for each dimension, the texts generated are semantically consistent, expressing a specific view but using different wordings, with a variety of expressions. Also, one interesting point of view regarding these generated texts is that they can be considered to be tweets written by users having a high score in that specific dimension, or holding the specific belief corresponding to the dimension.

The texts generated from the rest of 50 dimensions are included in the Supplementary Information file. Examining all the texts generated for each dimension, we find many novel interesting belief perspectives that haven't been found or explored in the original primal world beliefs work (Clifton et al. 2019). For example, dimension 15^{th} describes a view that the world is big, wide but cold and lonely, dimension

Dim.	Texts Generated	Dim.	Texts Generated
19^{th}	the world is ... ... an addiction, it destroys my life ... cruel and unforgiving and i will die alone ... an ugly place, it's killing me. ... a horrible place and life is unbearable. ... hell, and the only hope is to survive in it ... sick and it's unbearable to feel alone.	28^{th}	the world is ... ... changing but the revolution is timeless. ... rapidly changing into what we now know as the west. ... slowly but surely ending. ... changing faster than we know it. ... slowly but surely changing. ... rapidly changing into oblivion.
29^{th}	the world is ... ... so upside down, i swear. ... just plain crazy, it's like god is calling us. ... so f**king complicated, its just a matter of time ... so upside down, im just here to stay ... so crazy, so crazy now that i know it ... so crazy, it just seems so random	32^{th}	the world is ... ... truly in your favor ... yours for the taking. ... open to everybody. ... a better place now that u have been granted the blessing. ... truly yours, man. ... truly in the palm of your hands
46^{th}	the world is ... ... an illusion, your mind controlled by a stranger ... nothing but a playground and everyone plays ... nothing but a hologram sent to you by god ... an illusion, trapped within the brain ... invisible to the outside world, it's out there every bit as important as its in your brain ... like a mirror - only you see through it.	47^{th}	the world is ... ... a beautiful place right now, thanks god ... black and white and it is lonely watching ... a better place with you in it. ... a happier place now that the sun is out ... a brighter place with you in it. ... a beautiful place now that you awake

Table 1: Texts generated from selected latent belief dimensions.

50^{th} describes a view that the world belongs to me possessively, and that oneself has controlling power over the society surrounding them. These can be considered a good source of potential primals for psychologists to study further.

Quantitative Results

A key strength of our latent belief model LaBel is that we expect it to produce coherent and easy to interpret textual descriptions of the latent dimensions. As such, we want to assess the value of representing a latent dimension using generated sentences rather than using a collection of keywords. To evaluate this, we compare against descriptions of topics obtained through a standard topic modeling technique, LDA (Blei, Ng, and Jordan 2003). We used the scikit-learn Python library implementation of the LDA algorithm, with hyperparameters set as $learning\_method$ = "$online$", $batch\_size$ = 100, max_iter = 10, other hyperparameters were used as defaults.

Using the latent belief embeddings from our model, we use the following process to obtain representative keywords for each dimension: first, generate 100 sentences randomly from the interval $[\mu_i + 2 \times \sigma_i, \mu_i + 4 \times \sigma_i]$ of each interested dimension while keeping other dimensions at mean values, then get the most frequent 1-gram and 2-grams from these generated sentences and use them as representative keywords for that dimension.

We use the human-evaluating intrusion task proposed in Chang et al. (2009) to measure the coherence of the major beliefs found by these different approaches. With each approach, for each explored topic, we print 5 of its most representative items. For LaBel method with text generating version (LaBel-phrases), each representative item will be a sentence generated from the decoder. For LDA and LaBel with high frequencies words version (LaBel-words), representative items will be the most weighted or the most frequent words found by the algorithm. One of the 5 items will be replaced by an "intruder" from another topic. The performance of each method will be measured by the accuracy in choosing the correct intruders. In the extreme cases where the topic is not at all coherent, meaning all the representatives do not share any similarity semantically, then participants probably will choose the answer randomly among 5 items, and therefore the accuracy rate is just 20%, which is considered the baseline of the method. Four human evaluators completed the intrusion detection task and the accuracy for each method is averaged across the four.

Observing the results in Table 2, we see that each method provides incremental improvement over the previous with the largest gain coming from using generated phrases. Using the latent embeddings to either generate texts or keywords is slightly better than using keywords derived from LDA topic models. This might be because our method is based on transformer-based contextual embeddings which better capture semantics in context, while LDA mainly relies on the co-occurrence of words which in this case might be a problem since tweets are usually very short. We also see that the text generating approach to describe the dimension has a significantly higher score than using words as well as LDA. To be more specific, for "the world is...", LaBel-phrases' mean accuracy is 0.595, versus 0.25 of LDA (with $p < 0.001$ in difference significance test) and 0.295 of LaBel-words ($p < 0.001$). For "education is...", LaBel-phrases' mean accuracy is 0.375, versus 0.275 of LDA ($p = 0.065$) and 0.263

Method	Accuracy mean	Accuracy std
Baseline	0.2	-
Dataset 1: "the world is ..."		
LDA	0.250	0.091
LaBel-words	0.295	0.069
LaBel-phrases	**0.595**	0.017
Dataset 2: "education is ..."		
LDA	0.275	0.075
LaBel-words	0.263	0.114
LaBel-phrases	**0.375**	0.178
Dataset 3: "parents should ..."		
LDA	0.2	0.079
LaBel-words	0.175	0.075
LaBel-phrases	**0.4**	0.035

Table 2: Results for human evaluation intrusion task.

of LaBel-words ($p = 0.055$). For "parents should...", LaBel-phrases' mean accuracy is 0.4, versus 0.2 of LDA ($p < 0.05$) and 0.175 of LaBel-words ($p < 0.05$). This validates our hypothesis that having complete natural language phrases to convey the semantics is much more effective than sparse words alone.

Latent Beliefs Dimensions for Prediction

We also considered how well the factorized latent dimensions align with psychology research results as explored in (Clifton et al. 2019) by using these dimensions as features to predict the primals classes of tweets. Achieving high accuracy would help to show that the explored beliefs embeddings using our method align with psychologists' findings in the semantic space. We asked experts in human beliefs to annotate a set of 74 tweets to belong to a set of 6 classes of beliefs ("Safe", "Dangerous", "Enticing", "Dull", "Alive", "Mechanistic") which cluster into 2 primary classes ("Good" and "Bad"). We then use these tweets' latent dimensions as input features and their annotated classes as output labels to train a ridge classifier.

In this experiment, to demonstrate our approach's effectiveness on this problem, we also compare our method with other topic-modeling methods sharing similar properties to ours such as dealing with short texts, using neural embeddings features. To be specific, we considered the following methods: (1) LDA, (2) biterm for short text modeling (Yan et al. 2013) - a method that models the word-pair co-occurrence patterns in the whole corpus aiming to solve the problem of sparse word co-occurrence at document-level, (3) Contextualized Topic Models (CTM) (Bianchi, Terragni, and Hovy 2021; Bianchi et al. 2021) - a method adding BERT embeddings to improve Neural Topic Model, and (4) baseline - taking the mode of the training set as predictions. We used 10-folds cross-validation to evaluate each method, where we trained the model on 9 training folds and tested on the left-out testing fold. The results reported for each method are weighted precision, recall and F_1 scores, averaged from 10 testing folds. For each division, we chose the best ridge classifier model's hyperparameter α from searching within $[0.001, 0.01, 0.1, 1.0, 10]$. The CTM and biterm algorithm were implemented using *contextualized-topic-models* and *biterm* Python library respectively, which were developed by the original papers' authors. Hyperparameters for these algorithms were used as default values. Table 3 shows the results of the experiment.

Methods	precision (weighted)	recall (weighted)	F_1 (weighted)
2-classes			
baseline	0.145	0.367	0.206
LDA	0.540	0.497	0.466
biterm	0.360	0.406	0.335
CTM	0.668	0.567	0.552
LaBel	**0.763**	**0.708**	**0.704**
6-classes			
baseline	0.114	0.301	0.161
LDA	0.416	0.460	0.408
biterm	0.413	0.459	0.402
CTM	0.408	0.460	0.391
LaBel	**0.437**	**0.499**	**0.437**

Table 3: Results for predicting tweets' primals' classes from their latent beliefs embeddings

We observe that the features vectors from LaBel outperform those from other word co-occurrence-based methods, LDA and biterm, in both cases of 2-classes prediction and 6-classes prediction. This might be thanks to the utilization of contextualized embeddings in our method. Compared to CTM which also uses BERT embeddings, our model performs better, especially in the 2-classes task. Possibly because while CTM uses contextualized embeddings as input to a neural model to infer topic distributions and hence loses some information about the sentiments ("Good" or "Bad") from the tweets embeddings, our method, on the other hand, only applies NMF factorization to the original embeddings to get the features vectors and thus, preserve richer sentimental information.

Additionally, we also tested the robustness of our model regarding the choice of number of dimensions, in which we compressed the tweets embeddings into 25 and 100 dimensions (besides 50 dimensions). We found that, in these cases, our model still performs well and consistently outperforms other methods, similarly to when the number of dimensions is 50. Please see the Supporting Information files where we included the prediction results and qualitative generated texts for each of 25 and 100 dimensions experiments.

Generalizing to Other Datasets

To test the generalizability of our model, such as exploring other sets of beliefs, we conducted the human intruding evaluation on different beliefs set for *"education"* and *"parenting"*. To be more specific, we use phrases "education is"

Architectures	Validation Perplexity
Method 1 (1 transformation matrix for second last layer)	12.575
Method 2 (1 transformation matrix for all layers)	9.419
Method 3 (12 transformation matrices for all layers)	9.122

Table 4: Comparison of different modifications of GPT-2 to generate texts conditioned on a latent beliefs embedding

and "parents should", "parents need to" to extract tweets discussing these two concepts. Examples of the collected data are "education is the right of everyone", "parents need to learn that yelling doesn't work". We use the same pipeline as described in the previous section: extract BERT embeddings, apply NMF algorithm, and then build a GPT-2-based decoder to generate texts from each dimension. The only difference is the number of dimensions modeled is 20 instead of 50 due to the lower number of tweets in these two datasets (3,332 tweets for education topic and 2,005 tweets for parenting topic). We compared the three methods: LDA, LaBel-words, and LaBel-phrases on these 2 datasets and report the results in Table 2. The results again show that describing the explored topics with generated phrases is more understandable compared to using a list of representative words.

Latent Beliefs Embedding Transformations

Table 4 shows the comparison of 3 approaches for injecting the latent beliefs embedding information into the GPT-2 decoder: using the embedding only in the second last layer, using the embedding in all 11 layers but sharing a single transformation matrix, and using the embedding in all 11 layers but having layer-specific transformation matrices. We report the validation loss of the model trained on 80% of the training data and tested on 20% of validation data.

We find that using methods 2 and 3 (feeding the embedding to all 11 layers) outperforms method 1 (feeding to the second last layer only). The lower validation perplexity shows they can generalize and learn better from the data. Using layer-specific transformation matrices works better than using a single transformation matrix across all layers. Layer-specific transformations potentially help the model extract different information from the input vector that is at an appropriate level for each layer.

Applicability and Advantages of LaBel Model

The main application of our method is automating the process of exploring beliefs from a large dataset of texts, by generating natural texts with full contexts instead of discrete word lists to describe each latent dimension/topic with more precise semantics. This approach helps to rely less on experts' manual efforts as in previous psychology studies, significantly saving time, especially when analyzing large text corpus. This could also allow psychologists to more efficiently analyze fine-grained contexts of beliefs, such as beliefs from texts of different locations, political ideologies, or age groups, and beliefs regarding different subjects such as parenting, equality, and education. Examples of some belief research topics that might benefit include work on beliefs about exercise and it's effect on health outcomes (Crum and Langer 2007; Boswell et al. 2021). Other psychology research, e.g. (Beck et al. 1987; Hofmann et al. 2012) examines beliefs about the self (e.g. "I'm worthless"), which are known to impact depression and many other clinical outcomes, but the dimensionalities of these self-beliefs are not known and LaBel could be used to identify those dimensionalities. With the characteristics of being automatic and efficient on large datasets, our model can also be useful for gaining insights into contemporary and quickly-updated events from social media texts. LaBel could be used by psychologists and policy-makers to help define attitudes and beliefs about urgent policy-related topics where rampant misinformation (i.e., incorrect beliefs) impacts public health. For example, along the line of work of (Salali and Uysal 2020; Saied et al. 2021), LaBel could be used to identify the dimensionality of specific beliefs about the Covid-19 vaccines, so strategies for combating this misinformation can be determined.

Along with the above mentioned-potential applications of our model, LaBel also offers some advantages that make it a more reasonable choice than other methods in circumstances, especially when analyzing social media texts. First, while social media is one of the main platforms used to self-express and share opinions, it also encounters the problem of data privacy where users accidentally expose their personal information in posts or tweets. Our method helps preserve users' privacy by not showing the exact posts written by them, but instead generalizing the language across a large number of users regarding one belief. Second, after training the generator, psychologists can use the trained model to generate synthetic data corresponding to their interested dimensions. This could lead to applications with interactive/guided exploration of beliefs, such as regarding climate change or vacation, giving researchers a much more thorough qualitative understanding of beliefs and their contexts. Third, the ability to reduce human beliefs to a set number of dimensions allows theory checking and hypothesis building in the social sciences. The generation of texts overcomes the problem of word-based topic modeling regarding words taking on multiple meanings (e.g. consider 'crazy' in "the world is so crazy, upside down" versus "the world is crazy beautiful").

Besides the advantages of LaBel, we also want to discuss some limitations of our model and suggest potential solutions for future work. Regarding our decoder generating texts from latent dimensions, it works most effectively when our interested topics have the format of "X is ..." (e.g., "the world is ...", "parent should ..."), in which the topics lies at the beginning of a sentence. This is because most state-of-the-art texts generative models use this autoregressive learning mechanism in which texts are learned and generated from left to right. This learning mechanism is significantly less complex than generating texts from right to left,

or from prompt words in the middle of a sentence. Since we want to harness the power of state-of-the-art generative models such as GPT-2, we relied on this autoregression learning mechanism. However, our proposed method can still potentially be applied without significant modifications to other transformers-based non-autoregressive text generative models, such as from Liao, Jiang, and Liu (2020); Lawrence, Kotnis, and Niepert (2019); Stern et al. (2019), to avoid being limited to the "X is..." format, although the quality and coherence of the generated texts might be inferior due to the complexity of this non-autoregressive learning mechanism.

Conclusion

We proposed LaBel, a transformer-based model, to help automate the process of exploring people's major beliefs about the surrounding world, or "primals", by analyzing social media content and factorizing it into latent semantic dimensions. Our approach enables interpretation of these latent dimensions by generating example phrases from a modified version of a language model decoder. Viewing the approach as a modern alternative to topic modeling (which represents latent domains by their most prevalent words), we show that representing topics or latent dimensions by generated texts is more interpretable than the commonly used collections of keywords or n-grams as in traditional topic modeling. We also proposed a new approach for modifying the GPT-2 model to generate texts conditioned on a vector rather than special leading tokens or prompting texts as in other related works. As social media sites such as Facebook, Twitter, and Reddit become one of the leading platforms for people to express their thoughts, the proposed latent belief model offers potential use for psychologists as a scalable alternative for studying human beliefs on various topics, enabling a more data-driven approach that is less dependent on the experts' own belief biases. Further, we also see this approach as a foundation for others to build on for additional text generation tasks and methods conditioned on vectors.

Ethical Statement

Although we hope to have our proposed model applied for exploring public major beliefs and opinions of interest for research purposes, there is a chance that it would be used for negative or unethical applications. One example is to look for major hate speeches, discriminating threads aggregated on social media sites, which is currently an important concern on these platforms. Although it is difficult to restrict what readers can do with our proposed model, we do suggest that readers be responsible when citing and taking advantage of our works to use them for positive impact research only. Regarding the Twitter data provided along with this study, we have preprocessed them using deterministic codes to remove identifiable information of participants. We suggest readers not use our data for unethical reasons and suggest the good practice of preserving participants' personal information when collecting data for training with our proposed model.

References

Adelani, D. I.; Mai, H.; Fang, F.; Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. arXiv:1907.09177.

Beck, A. T.; Rush, A. J.; Shaw, B. F.; and Emery, G. 1987. Cognitive therapy of depression. *The Australian and New Zealand journal of psychiatry*.

Bianchi, F.; Terragni, S.; and Hovy, D. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 759–766. Online: Association for Computational Linguistics.

Bianchi, F.; Terragni, S.; Hovy, D.; Nozza, D.; and Fersini, E. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1676–1683. Online: Association for Computational Linguistics.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*.

Boswell, M. A.; Evans, K. M.; Zion, S. R.; Boles, D. Z.; Hicks, J. L.; Delp, S. L.; and Crum, A. J. 2021. Mindsets Predict Physical Activity and Relate to Chosen Management Strategies in Individuals with Knee Osteoarthritis. *medRxiv*.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 10–21. Berlin, Germany: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; and Blei, D. M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, 288–296. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781615679119.

Clifton, J. D. W. 2020. Testing If Primal World Beliefs Reflect Experiences—Or at Least Some Experiences Identified ad hoc. *Frontiers in Psychology*, 11: 1145.

Clifton, J. D. W.; Baker, J. D.; Park, C. L.; Yaden, D. B.; Clifton, A. B. W.; Terni, P.; Miller, J. L.; Zeng, G.; Giorgi, S.; Schwartz, H. A.; and Seligman, M. E. P. 2019. Primal world beliefs. *Psychological Assessment*, 31: 82–99.

Coppersmith, G.; Dredze, M.; Harman, C.; and Hollingshead, K. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1–10.

Crum, A. J.; and Langer, E. J. 2007. Mind-set matters: exercise and the placebo effect. *Psychological science*, 18: 165–171.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dieng, A. B.; Ruiz, F. J. R.; and Blei, D. M. 2020. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8: 439–453.

Ganesan, A. V.; Matero, M.; Ravula, A. R.; Vu, H.; and Schwartz, H. A. 2021. Empirical Evaluation of Pre-trained Transformers for Human-Level NLP: The Role of Sample Size and Dimensionality. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, 4515–4532. Association for Computational Linguistics.

Hofmann, S. G.; Asnaani, A.; Vonk, I. J.; Sawyer, A. T.; ; and Fang, A. 2012. The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognitive therapy and research*, 36: 427–440.

Hofmann, T. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, 50–57. New York, NY, USA: Association for Computing Machinery. ISBN 1581130961.

Kulkarni, V.; Kern, M. L.; Stillwell, D.; Kosinski, M.; Matz, S.; Ungar, L.; Skiena, S.; and Schwartz, H. A. 2018. Latent human traits in the language of social media: An open-vocabulary approach. *PloS one*, 13(11).

Lawrence, C.; Kotnis, B.; and Niepert, M. 2019. Attending to Future Tokens for Bidirectional Sequence Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1–10. Hong Kong, China: Association for Computational Linguistics.

Lee, D. D.; and Seung, H. S. 2000. Algorithms for Non-Negative Matrix Factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, 535–541. Cambridge, MA, USA: MIT Press.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Liao, Y.; Jiang, X.; and Liu, Q. 2020. Probabilistically Masked Language Model Capable of Autoregressive Generation in Arbitrary Word Order. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 263–274. Online: Association for Computational Linguistics.

Matero, M.; Idnani, A.; Son, Y.; Giorgi, S.; Vu, H.; Zamani, M.; Limbachiya, P.; Guntuku, S. C.; and Schwartz, H. A. 2019. Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Pilault, J.; Li, R.; Subramanian, S.; and Pal, C. 2020. On Extractive and Abstractive Neural Document Summarization with Transformer Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9308–9319. Online: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Saied, S. M.; Saied, E. M.; Kabbash, I. A.; and Abdo, S. A. E. 2021. Vaccine hesitancy: Beliefs and barriers associated with COVID-19 vaccination among Egyptian medical students. 93: 4280–4291.

Salali, G. D.; and Uysal, M. S. 2020. COVID-19 vaccine hesitancy is associated with beliefs on the origin of the novel coronavirus in the UK and Turkey. *Psychological medicine*.

Santhanam, S.; and Shaikh, S. 2019. Emotional Neural Language Generation Grounded in Situational Contexts. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, 22–27. Tokyo, Japan: Association for Computational Linguistics.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PloS one*.

Stahlmann, A. G.; Hofmann, J.; Ruch, W.; Heintz, S.; and Clifton, J. D. 2020. The higher-order structure of primal world beliefs in German-speaking countries: Adaptation and initial validation of the German Primals Inventory (PI-66-G). *Personality and Individual Differences*, 163: 110054.

Stern, M.; Chan, W.; Kiros, J.; and Uszkoreit, J. 2019. Insertion Transformer: Flexible Sequence Generation via Insertion Operations. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 5976–5985. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Vu, H.; Abdurahman, S.; Bhatia, S.; and Ungar, L. 2020. Predicting Responses to Psychological Questionnaires from Participants' Social Media Posts and Question Text Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1512–1524. Online: Association for Computational Linguistics.

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. arXiv:1901.08149.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A Biterm Topic Model for Short Texts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 1445–1456. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320351.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278. Online: Association for Computational Linguistics.