Automated Emotional Valence Prediction in Mental Health Text via Deep Transfer Learning

Benjamin Shickel*§, Martin Heesacker†§, Sherry Benton¶, Parisa Rashidi‡§
*Department of Computer and Information Science and Engineering, †Department of Psychology, ‡Department of Biomedical Engineering, §University of Florida, ¶TAO Connect

***Abstract*—Sentiment analysis is a well-researched field of machine learning and natural language processing generally concerned with determining the degree of positive or negative polarity in free text. Traditionally, such methods have focused on analyzing user opinions directed towards external entities such as products, news, or movies. However, less attention has been paid towards understanding the sentiment of human emotion in the form of internalized thoughts and expressions of self-reflection. Given the rise of public social media platforms and private online therapy services, the opportunity for designing accurate tools to quantify emotional states in is at an all-time high. Based upon psychological research, in this work we propose a new type of sentiment analysis task using a two-dimensional valence scheme with four sentiment categories: positive, negative, both positive and negative, and neither positive nor negative. This work details the collection of a novel annotated dataset of real-world mental health therapy logs and compares several machine learning methodologies for the accurate classification of emotional valence. We found superior performance using deep transfer learning approaches, in particular using the recent breakthrough method of BERT. We argue that representing emotional sentiment on decoupled valence axes is an appropriate modification of traditional sentiment analysis for mental health tasks and that modern transfer learning approaches should become an essential component of automated mental health frameworks, where labeled data is often scarce.**

I. Introduction

Sentiment analysis, the task referring to the automatic determination of user opinion from text, has received increased attention in the past decade [1]–[4]. Much of the success of sentiment analysis techniques can be attributed to the rise of social media platforms, where millions of users share their opinions on a wide variety of subjects. The majority of sentiment analysis methods are aimed at aggregating opinions towards entities like movies, people, products, or companies. We refer to this well-known research area as external sentiment analysis, in which sentiment and textual polarity is calculated with respect to a specific external entity. In contrast, we define internal sentiment analysis as the study of the polarity of user text with respect to themselves, primarily concerned with statements of emotion and mental health [5]. In this paper, we strictly focus on internal sentiment analysis, specifically with the valence prediction of private journals in a mental health therapy setting. Our work partly aligns with previous research regarding emotion detection in text [6]–[8], a subtask of the field of affective computing and analysis, but unlike previous work, we focus on the expansion of valence categories in a mental health setting.

One useful application of automated internal sentiment analysis is online mental health therapy services [9]–[12]. These programs provide education components such as cognitive behavioral therapy (CBT), which enable patients to identify and change unhelpful thought patterns. Additionally, users often document their daily thoughts and feelings in an online journal to help with self-directed therapy strategies. Providing feedback to patients in this process can increase their ability to accurately identify positive and negative thoughts and improve therapy outcomes.

Until now, ongoing feedback for self-directed CBT-type programs has been difficult to provide. An accurate and validated system for automatically categorizing polarity of user text has obvious benefits, such as flagging text which may be an early warning for suicide risk, providing a positive and always-available feedback for patients with distorted thinking, or simply providing enhanced and more fine-grained analysis of overall patient mental well-being for therapists.

Traditional sentiment analysis involves detecting whether a given text fragment is subjective or objective, and in the case of subjectivity, classifies the text as either positive or negative. We take this analysis a step further for mental health polarity, where rather than framing the subjectivity identification task as a binary classification between positivity ($\mathcal{P}$) and negativity ($\mathcal{N}$), we introduce two additional classes of polarity: both positive and negative ($\mathcal{PN}^+$), and neither positive nor negative ($\mathcal{PN}^-$). We made this decision based on psychological research which suggests emotions cannot be represented on a single axis of valence [13]–[17]. Using our new annotation scheme, text previously classified as neutral would fall into either of the two augmented classes.

Unfortunately, publicly-available mental health datasets suitable for machine learning-based internal sentiment analysis are few and far between. However, large amounts of social media text have become available in recent years, and several studies have examined traditional sentiment analysis in the context of social media platforms such as Twitter [18]–[20]. Given social media users' tendency towards self-expression, we hypothesized that the social media domain is quite similar to the mental health domain with regards to language modeling and classification and can be used to help train models for mental health analysis.

In this work, we explore the application of a machine learning technique known as transfer learning, an approach involving training models on one domain (e.g. social media text) and fine-tuning them on another target domain (e.g.

2471-7819/20/$31.00 ©2020 IEEE
DOI 10.1109/BIBE50027.2020.00051

mental health text). We specifically focus on transfer learning in the context of deep learning, a subfield of machine learning that is designed to automatically learn ideal representations of raw data without human supervision or manual feature engineering. In recent years, deep learning techniques have yielded state-of-the-art performance in the computer vision and natural language processing (NLP) domains. Success in NLP applications is due largely in part to these nonlinear models' ability to model language from raw characters or words.

In our first transfer learning setting, we train deep learning models on a large dataset of annotated text from the social media platform Twitter and transfer the underlying language model and learned transformations to the task of predicting the valence of mental health journal text. We demonstrate that language models built from social media can be successfully transferred to the mental health domain for improved performance for emotional valence prediction and argue that social media is a valuable source of data for text-based mental health applications, a domain in which labeled data is notoriously scarce.

In our second more general transfer learning experiment, we apply the recent domain-agnostic technique of BERT (Bidirectional Encoder Representations from Transformers) [21], which has achieved state-of-the-art results in a variety of NLP tasks. In particular, we use a BERT model pretrained on a combination of the BookCorpus dataset [22] and the entirety of the English Wikipedia for a more domain-agnostic experiment setting.

II. Methods

In this section, we provide implementation details, data description, and experimentation setup and methodology. We begin by describing the process by which we transfer knowledge from the social media domain to the mental health domain.

A. Transfer Learning

Generally, for machine learning applications when labeled data in the target domain is abundant, single-domain models are preferable due to their ability to recognize and generalize domain-specific patterns in the input. If labeled data is scarce, however, models have the tendency to overfit the training data by incorrectly attributing significance to small input variations and noise, and inevitably fail to generalize well. This is especially true for more complex models, including most deep learning techniques.

The primary idea behind transfer learning involves training a complex model on a source domain with large labeled dataset, and then transferring some or all of what the model has learned to a target domain. Much recent work on NLP transfer learning has focused on highly successful deep learning techniques [21], [23], [24]. The transfer learning process is especially helpful when training deep learning models, which inherently learn a characteristic domain representation from raw data. If the source and target domains are similar, this learned representation can be utilized for similar tasks in other domains with data scarcity. For image processing, this representation could be the recognition of edges or shapes; for natural language processing, generalized notions of syntax and semantics can be transferred.

Our target domain includes individuals' text-based emotional self-expressions and reflections from a cognitive-behavioral perspective, obtained through an online mental health therapy service. Unfortunately, large public datasets relating to this field do not exist in any meaningful quantity, which is our prime motivation for the application of deep transfer learning. We hypothesized that the social media domain, specifically a large number of public tweets from Twitter, would be similar enough to transfer knowledge of content, style, and structure to the personalized thoughts contained in mental health monitoring logs.

Given recent advances in deep learning for a variety of natural language processing tasks, we explore the application of recurrent neural networks for transferring social media knowledge to the mental health domain. While a full review of deep learning [25] is beyond the scope of this paper, we provide a general overview of our particular model in the following section.

B. RNN Transfer Learning

The architecture we employ for our first transfer learning experiment (Figure 1) is the recurrent neural network (RNN), a class of deep learning architectures especially suited for processing sequential data [26]. The RNN is a type of feedforward neural network that incorporates the notion of sequential memory, where the final sequence representation is a combination of the hidden representations from all input time steps. For RNNs operating as supervised classifiers, such as ours, the entire sequence representation is fed to a fully-connected output layer of size C, where C is the number of classes available for prediction, upon which a $softmax$ activation function generates a prediction probability distribution over the available classes. In this study, we use an RNN variant that uses gated recurrent units (GRU) [27], a system of gates that controls information flow within the network that has been shown to improve RNN performance by capturing longer-term dependencies and preventing exploding gradients. Additionally, we include a word-level attention mechanism based on the work of Yang, et al. [28] that learns to focus on meaningful input words to provide a measure of natural interpretability to our final model evaluation.

In our RNN transfer learning experiments, we first train a GRU-RNN on a large, labeled corpus containing two sentiment classes (positive and negative). We then create a new GRU-RNN designed to predict four emotional valence classes, and initialize the word embeddings and model weights to the values found from the first step. This new model is then fine-tuned on our mental health dataset.

C. BERT Transfer Learning

In the NLP domain, deep transfer approaches involve the training of unsupervised language models from large unlabeled corpora which can be subsequently fine-tuned for a variety of

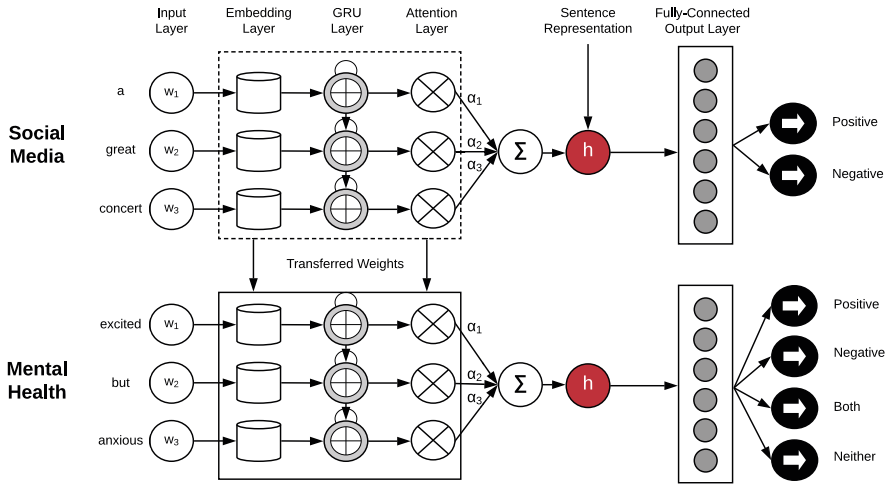

Fig. 1. RNN framework for transferring knowledge from the social media domain to the mental health domain. After the binary classification model is trained solely on social media, we fine-tune the model to the new task with the four-way emotional valence classification objective.

downstream tasks including sequence classification, question answering, summarization, and part-of-speech tagging. While similar in spirit to the weight-copy and fine-tuning procedure described in the previous section, these newer techniques differ in that they no longer require labeled pretraining corpora, instead building generalized language models from any collection of text, including massive datasets such as the entirety of Wikipedia [21]. The vastness of available training data enables more complex model designs, which once fine-tuned, have been proven to work well for even small datasets.

For our second transfer learning experiment, we employed a pretrained BERT model due to its reported state-of-the-art performance on a wide variety of NLP tasks [21]. At its core, BERT is a Transformer model, a type of deep learning architecture suited for processing sequential data without any explicit recurrence, unlike traditional RNN-based sequential models such as the LSTM or GRU. A full description of the Transformer network is beyond the scope of this study; a comprehensive overview can be found in Vaswani, et al [29].

Briefly, what sets BERT apart from other recent unsupervised NLP transfer methods is in its novel pretraining method. While training the language model, pairs of sentences are fed into the network simultaneously, where 50% of the time the second sentence is the actual sentence following the first in the original corpus, and the remainder of the time the second sentence is another random sentence from the corpus. Additionally, in each input sentence, 15% of the tokens are modified to either be replaced with a masking token, random word, or keep unchanged. Sentence tokens are embedded by the summation of token embeddings, sentence embeddings, and positional embeddings. The training loss involves the summation of the word-prediction from masked tokens and the prediction of whether the second sentence is the true subsequent sentence.

BERT is designed to facilitate easy application to any type of downstream NLP task, such as the sequence classification involved in emotional valence prediction. Given a pretrained BERT model, a final classification layer is appended to the architecture, and the entire model is fine-tuned with labeled data in the target domain. An overview of adapting BERT for mental health valence prediction is shown in Figure 2.

Our use of the BERT model, pretrained on a large unlabeled dataset including the entirety of Wikipedia and BookCorpus, represents a domain-agnostic approach to transfer learning, in contrast to the RNN-based transfer learning where knowledge from a single domain (social media) is transferred to a new emotional valence prediction model.

III. Experiments

Our target dataset includes 3,872 IRB-approved responses retrospectively collected from monitoring logs submitted by users as part of the TAO Connect[1] online therapy program. Textual responses were annotated by three trained psychology graduates, directed under the supervision of departmental faculty, to provide valence labels for these responses. The annotators were tasks with identifying emotional effect in the responses, labeling each as either positive ($\mathcal{P}$), negative ($\mathcal{N}$), both positive and negative ($\mathcal{PN}^+$), or neither positive nor negative ($\mathcal{PN}^-$). Final labels were assigned via a majority voting scheme, and in the case of a three-way tie, the response was discarded.

Of the 3,872 responses, 63.5% were assigned the $\mathcal{N}$ label (10.4% $\mathcal{P}$, 14.3% $\mathcal{PN}^-$, and 11.7% $\mathcal{PN}^+$). The large class skew is a side effect of selection bias in our dataset - most psychotherapy seekers exhibit negative thinking, which is one major reason people seek therapy. We preprocessed the target domain text using the same methods as the source domain.

The social media dataset used in our related domain transfer task comes from Sentiment140[2], a public dataset designed for traditional sentiment analysis tasks. The data contains 1.6 million tweets annotated as positive or negative by an automatic process using the presence of specific emoticons. Our

[1]http://taoconnect.org
[2]http://sentiment140.com

preprocessing steps included lowercasing, hyperlink removal, user mention removal, and punctuation stripping.

A. Model comparison

Our experiments are designed to evaluate the effectiveness of deep transfer learning in comparison with baseline models trained in only the target domain. On the four-way emotional valence prediction task. We compare two baseline methods with two transfer learning approaches. The models included in our comparison are summarized below.

- **Single-domain conventional classifier:** For this baseline model, we use the traditional NLP technique of vectorizing the mental health text using a bag-of-words approach, normalizing by term frequency-inverse document frequency (tf-idf). A linear logistic regression classifier is used to predict emotional valence from the vectorized text.
- **Single-domain RNN:** The second baseline model is a bidirectional recurrent neural network (RNN) with gated recurrent units (GRU) and a word-level attention mechanism. All network weights are initialized randomly and learned only from the target domain as training progresses.
- **RNN with supervised domain transfer:** We first train a bidirectional RNN with gated recurrent units and word-level attention mechanism on a large social media dataset annotated for positive and negative polarity. We then fine-tune the model on our novel mental health dataset after replacing the original binary classification output layer with a four-way linear layer representing the four emotional valence categories. This model transfers knowledge gained from a related domain in the form of pretrained word embeddings, recurrent weights, and attention weights.
- **Pre-trained BERT model with fine-tuning:** Our final model incorporates a BERT framework pre-trained on the BookCorpus and English Wikipedia datasets. We append a final classification layer to the overall pretraining architecture and jointly train the entire model for our emotional valence prediction task in a fine-tuning procedure.

For consistency, we set all RNN parameters equal to the same values and do not perform hyperparameter search. Specifically, we use 50% dropout on word vectors, 64 bidirectional GRU units, tanh activation, 50% dropout on the RNN output, and use an Adam optimizer with learning rate of 0.01. The pretrained BERT model (BERT-Base, Uncased)[3] consists of 12 hidden layers with 768 hidden units and 12 multi-attention heads. All methods involving deep learning were performed with early stopping based on three successive epochs with no improvement in macro F1 score.

For all models described above, we report 5-fold cross-validation results on the mental health dataset. Specifically, we first split the mental health dataset into five non-overlapping folds of 80% training data and 20% testing data. For each model, predictions were made on each of the five test folds after using the train fold for constructing the model. Performance metrics were calculated from the predictions of the entire mental health dataset by concatenating each of the test folds. For transfer learning settings, the train fold is used for fine-tuning the pretrained model. Each fold is stratified in a balanced manner to reflect the global distribution of valence classes.

[3]https://github.com/google-research/bert

TABLE I
EMOTIONAL VALENCE CLASSIFICATION RESULTS.

		Single domain		Transfer learning	
Label	**Metric**	**Logistic**	**RNN**	**RNN**	**BERT**
$\mathcal{P}$	Precision	0.58	0.46	0.61	**0.72**
	Recall	0.44	0.53	0.55	**0.65**
	F1	0.50	0.49	0.58	**0.68**
$\mathcal{N}$	Precision	0.82	0.86	**0.92**	**0.92**
	Recall	0.91	0.81	0.84	**0.94**
	F1	0.86	0.83	0.88	**0.93**
$\mathcal{PN}^+$	Precision	0.50	0.42	0.47	**0.62**
	Recall	0.39	0.48	0.62	**0.64**
	F1	0.44	0.44	0.54	**0.63**
$\mathcal{PN}^-$	Precision	0.69	0.63	0.66	**0.79**
	Recall	0.59	0.67	**0.79**	0.76
	F1	0.64	0.65	0.72	**0.78**
All	Accuracy	0.75	0.72	0.78	**0.85**
	Macro F1	0.61	0.60	0.68	**0.76**
	Weighted F1	0.74	0.72	0.79	**0.85**

IV. RESULTS

Full valence classification results are shown in Table I, where BERT was the best performing model from the perspective of both label-wise metrics (6-17% higher F1 score, up to 24% higher precision, up to 18% higher recall) and global dataset-wide metrics (9% higher accuracy, 12% higher macro F1 score, and 8% higher weighted F1 score).

For single domain models trained and tested only on the mental health dataset and with no external pretraining or knowledge transfer, the logistic regression classifier outperformed the RNN. However, by incorporating social media knowledge transfer by means of fine-tuning the RNN trained on the Twitter dataset, performance increased for all metrics. Neither RNN model outperformed the general-purpose BERT model, which yielded superior performance over all other models by a significant margin.

V. DISCUSSION

Both deep learning models that take advantage of transfer learning from external sources exhibit substantial increases over the linear baseline in recall for the augmented valence classes $\mathcal{PN}^+$ and $\mathcal{PN}^-$, suggesting that the subtle cues differentiating these classes from the traditional "neutral" class are missed by simple linear models, and that learning traits of social media language use can be useful for mental health tasks.

The reduced performance of the RNN without any knowledge transfer can be explained by the small size of the dataset,

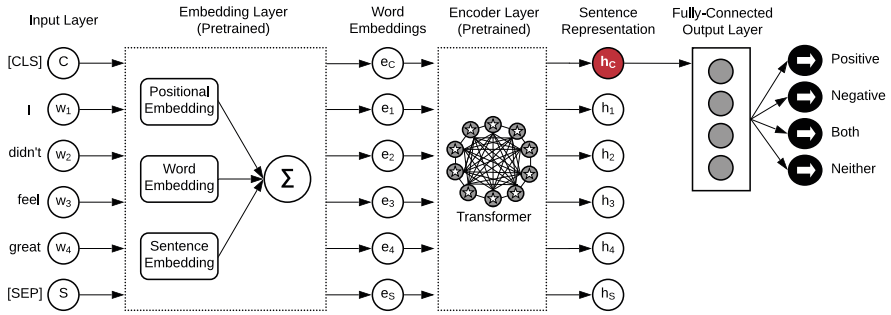

Fig. 2. Overview of transferring knowledge from a pretrained, domain-agnostic BERT model to the mental health domain. *[CLS]* is a special token which takes on the representation of the entire input sentence for use in the final classification layer. *[SEP]* is another special token indicating the boundary between two sentences.

which resulted in large amounts of overfitting. Pretraining the RNN on social media text had a positive impact on all classification metrics but was still outperformed by the more general-purpose language model of BERT.

A software tool incorporating this study's prediction models has important implications for computer-assisted, computerized psychotherapy, and other therapy-related applications in which patients submit narrative text. This tool can be used to identify circumstances where negative affect in patient narratives is high, suggesting elevated levels of distress, depression, or anxiety. It can also be used to detect cases in which positive and negative affect are both high in narrative text, which may be indicators of ambivalence, which is often associated with conflict-related distress. In addition to detecting static high levels of negative affect or high levels of both negative and positive affect, this tool can be used to detect dynamic changes across time. Declines in negative affect or declines in simultaneously high levels of positive and negative affect in narrative responses is an indication of symptom reduction and perhaps a sign of treatment effectiveness.

In contrast, increases in negative affect or increases in simultaneously high levels of positive and negative affect in narrative responses is an indication of worsening symptoms and perhaps a sign of additional trauma and/or treatment ineffectiveness. These signs of worsening symptoms can signal the need to review and perhaps change treatment, to deal with recent additional trauma and/or to deal with adverse responses to specific interventions. Unchanging or minimally changing elevated levels of negative affect in written narratives across time indicate a lack of treatment effectiveness or a plateau in treatment, either of which may require making different treatment decisions. Cyclical oscillations in levels of negative affect or oscillations between high negative and high positive affect across time suggest the possibility of bi-polar disorder, cyclothymia, borderline personality disorder, and/or an interpersonal or social context fraught with cyclical affective instability. Thus, this tool can be useful in evaluating treatment effectiveness, evaluating when to change treatment strategies, detecting patient social stressors, and validating or invalidating initial clinical diagnoses. In addition, the tool can be used as an indication of treatment completion. When negative affect in narratives declines and remains low, clinicians could explore whether it is time to terminate therapy.

Not only can this tool be useful in making the general conclusions just discussed, but it can also be used to provide immediate feedback to patients, whether automated feedback or feedback from the therapist. Comments that reflect empathy for the patient's current affective state can produced by analyzing for dynamic changes or stability in the affect reflected in narrative responses. Patients whose affect is predominantly negative across time can be encouraged to hold on and prompted to accept that change often takes more time than one would like. Patients whose negative affect is declining or whose positive affect is increasing, or ideally both, can be prompted to savor and appreciate the improvement, and to reflect on the factors associated with that improvement. Those with cyclical patterns can be encouraged to be aware of and reflect on the oscillation and perhaps to explore factors contributing to the oscillation, and ways to cope with it.

The tool can also be used to help draw patients' attention to the affective conditions reflected in their narratives and to cthe patterns reflected in these narratives across time. This attention-drawing process can help patients access aspects of themselves that might otherwise be outside of their awareness. For example, patients who have struggled with long-term mental health concerns may not consciously realize they are improving until they receive feedback about the improving affective patterns reflected in the tool's analysis of their narratives. Another example is a patient who struggles with regular affective oscillations but is unaware of the pattern across time and therefore limited in the ability to analyze, prevent, plan for, and cope with these cycles. A third example is that without feedback from the affect evaluation tool, patients and their therapists may not be aware of their deteriorating affective state until their situation is dangerous or catastrophic. A final example of patient self-awareness that can be enhanced by this tool involves premature termination. When patients still experience acute distress or significant clinical symptoms remain, receiving feedback that their narrative feedback is showing signs of improvement may encourage them to remain in treatment when they might otherwise drop out.

Yet another benefit of using this tool is the believability of the feedback. In situations in which patients might be skeptical of the accuracy of therapist feedback or reactive to the fact

that a healthcare provider is providing feedback, those same patients might place their trust in the more objective feedback from the machine learning tool. In fact, there is evidence that ostensibly objective personality assessment feedback is widely accepted, especially if it is worded somewhat positively and somewhat generally [30] though future studies will need to be conducted to assess whether feedback from this tool is widely accepted by patients.

The tool can also be used by therapists and therapy developers to assess the effectiveness of specific computerized modules or specific interventions, and can be used to assess whether these modules and interventions are differentially effective for different types of patients, for different types of presenting concerns, at different times in the therapy process, and whether different sequences of modules or interventions produce different affective responses in patient narratives. Finally, the tool can be used to assess whether particular identifiable affective patterns emerge over time from patients diagnosed with different disorders.

Our study has several limitations. First, it only focuses on the particular task of emotional valence prediction. Additionally, our target domain data came from a single institution, which resulted in a particular class skew that might not reflect global populations. In future work, we plan to explore a variety of mental health tasks and applications such as suicide prevention and identification of distorted thought patterns. Additionally, we plan to explore the inverted transfer setting, where private mental health journals may prove useful for predictive tasks in the public social media domain.

ACKNOWLEDGMENT

Research supported by NSF-IIP 1631871 from the National Science Foundation (NSF), Division of Industrial Innovation and Partnerships (IIP). We thank TAO Connect for access and assistance with retrieving online therapy logs. The Titan X used for this work was donated by the NVIDIA Corporation through the GPU Grant Program.

REFERENCES

[1] B. Liu, *Sentiment Analysis and Subjectivity*, 2nd ed., 2010.

[2] ——, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[3] S.-M. Kim, E. Hovy, S.-M. Kim, E. Hovy, and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th International Conference on Computational Linguistics*, Morristown, NJ, USA, 2004, pp. 1367–1373.

[4] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[5] B. Shickel, M. Heesacker, S. Benton, A. Ebadi, P. Nickerson, and P. Rashidi, "Self-Reflective Sentiment Analysis," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 23–32.

[6] H. Lee, Y. S. Choi, S. Lee, and I. P. Park, "Towards unobtrusive emotion recognition for affective social communication," *2012 IEEE Consumer Communications and Networking Conference, CCNC'2012*, pp. 260–264, 2012.

[7] E.-J. Lee and S. Y. Shin, "Are They Talking to Me? Cognitive and Affective Effects of Interactivity in Politicians' Twitter Communication," *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 10, pp. 515–520, 2012.

[8] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 2379–2392, 2013.

[9] B. M. Shaw, G. Lee, and S. Benton, "Work Smarter, Not Harder: Expanding the Treatment Capacity of a University Counseling Center Using Therapist-Assisted Online Treatment for Anxiety," in *Career Paths in Telemental Health*, M. M. Maheu, K. P. Drude, and S. D. Wright, Eds. Cham: Springer International Publishing, 2017, pp. 197–204.

[10] S. A. Benton, M. Heesacker, S. J. Snowden, and G. Lee, "Therapist-assisted, online (TAO) intervention for anxiety in college students: TAO outperformed treatment as usual." *Professional Psychology: Research and Practice*, vol. 47, no. 5, pp. 363–371, 2016.

[11] M. F. Travers and S. A. Benton, "The Acceptability of Therapist-Assisted, Internet-Delivered Treatment for College Students," *Journal of College Student Psychotherapy*, vol. 28, no. 1, pp. 35–46, 2014.

[12] J. Ruwaard, A. Lange, B. Schrieken, C. V. Dolan, and P. Emmelkamp, "The effectiveness of online cognitive behavioral treatment in routine clinical practice," *PLoS ONE*, vol. 7, no. 7, 2012.

[13] N. M. Bradburn, *The structure of psychological well-being*. Chicago: Aldine Publishing Company, 1969.

[14] P. Warr, J. Barter, and G. Brownbridge, "On the Independence of Positive and Negative Affect," *Journal of Personality and Social Psychology*, vol. 44, no. 3, pp. 644–651, 1983.

[15] E. Diener, R. J. Larsen, S. Levine, and R. a. Emmons, "Intensity and frequency: dimensions underlying positive and negative affect." *Journal of personality and social psychology*, vol. 48, no. 5, pp. 1253–1265, 1985.

[16] D. Watson, L. A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect - the Panas Scales," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1988.

[17] J. W. Pennebaker, "Putting stress into words: Health, linguistic, and therapeutic implications," *Behaviour Research and Therapy*, vol. 31, no. 6, pp. 539–548, 1993.

[18] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, 2011, pp. 538–541.

[19] A. Pak and P. Paroubek, "Twitter for Sentiment Analysis: When Language Resources are Not Available," *2011 22nd International Workshop on Database and Expert Systems Applications*, pp. 111–115, 2011.

[20] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Languages in Social Media*, 2011, pp. 30–38.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[22] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 19–27, 2015.

[23] M. E. Peters, M. Newmann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv*, 2018.

[24] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 328–339.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[26] D. P. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. John Wiley & Sons, 2001.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *IEEE International Conference on Rehabilitation Robotics*, vol. 2015-Septe, pp. 119–124, 2015.

[28] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, jun 2016, pp. 1480–1489.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[30] C. R. Snyder, R. J. Shenkel, and C. R. Lowery, "Acceptance of Personality Interpretations : The " Barnum Effect " and Beyond," *Journal of consulting and clinical psychology*, vol. 45, no. 1, pp. 104–114, 1977.