Using Big Data Methods to Identify Conceptual Frameworks

Robert Thorstad (rthorst@emory.edu)
Phillip Wolff (pwolff@emory.edu)
Emory University Department of Psychology

Abstract

Conceptual frameworks such as religion or politics may play a pervasive role in people's interpretation of experience, but the empirical evidence for such effects is limited. To the extent that conceptual frameworks are real, they should have a pervasive impact on how people talk about the world. Such an influence may be detected in people's everyday language. In a series of studies, text from the social media platform Reddit was used to train machine learning classifiers to identify people's association with a particular religion or mental disorder. Impressively, classifiers trained on text focusing on religion and mental disorders could be used to identify people's association with a particular religion or mental disorder even when the text was not explicitly about these topics, such as when it was about buying a car or playing tennis. Not only could the classifiers predict people's religion or mental illness in the present, they could also do so prospectively, indicating that people's everyday language gives away information about the kinds of conceptual frameworks they may hold in the future. An analysis of the features learned by the classifier suggested that they learned features with high face validity for the underlying conceptual framework. Together, the results provide evidence for the existence of conceptual frameworks by virtue of the imprint they leave across a wide range of language contexts.

Keywords: conceptual framework; big data; machine learning; social media.

Introduction

Intuition suggests that people may use conceptual frameworks to help interpret experience. They may use their general views about religion and politics, for example, to make sense of events and predict the future. Potential evidence for the existence of these frameworks is now coming from an unexpected source. Conceptual frameworks, if they exist, should have an effect across a wide range of contexts. One's religion or politics should, for example, have an impact not only on how one talks about religion and politics, but also on how one talks about incidental topics such as weekend activities or bad breakups. Conceptual frameworks, if real, should impact people's everyday talk. Recent studies applying big data techniques to social media are finding such effects.

The general strategy in these studies has been to train machine learning models that take people's everyday language or behavior as input and output a prediction about the probability of a particular mental perspective. The findings show, for example, that people's online language gives away information about their mental health (for review, see Guntuku et al, 2017) and personality (Youyou, Kosinski, & Stillwell, 2015). Such work has also found that everyday non-linguistic behaviors such as liking a page on Facebook (Kosinski, Stillwell, & Graepel, 2013) or choosing a profile picture (Liu, Petro, Samani, Moghaddam, & Ungar, 2016) predict people's demographics and personality. The success of these classifiers suggests that people cannot help but give away their general perspective in their everyday language and behavior.

The results so far potentially support the discovery of conceptual frameworks, but not necessarily. When trained on people's everyday language, the learning algorithms may not be picking up on people's overarching conceptual frameworks, but instead discovering different ways to solve the same kind of classification problem. Such a possibility is suggested in the results of several studies investigating conceptual frameworks. For example, Schwartz et al (2013) found that women are more likely to say *sooo* while men are more likely to use profanity, and Kosinski, Stillwell, & Grapel (2013) found that liking Britney Spears and the TV show Desperate Housewives are predictors of homosexuality. Such features seem quite distant from the overarching conceptual frameworks associated with sex and gender.

It should be possible to determine whether a classification model has learned a general conceptual framework, or has instead learned a specific set of predictors that solve a classification problem without necessarily learning a general perspective. Establishing whether a general perspective has been acquired can be achieved in two steps. First, a classifier can be trained on text from a particular context. Second, the performance of the classifier can be tested on a very different context. Such a strategy would be able to show whether the signal captured in the classifier in the first context represents a general perspective that could be applied to text from a different context. Because the context has been changed while the model has not, such generalization would be evidence for a consistent mental framework. The proposed strategy is very close to ordinary cross-validation. The key difference is that in cross-validation the training and tests sets come from the same set of contexts, while in the proposed test procedure, the training and test sets come from very different language contexts. The proposed strategy also has similarities to cross-domain text classification in machine learning. However, cross-domain classification usually focuses on improving transfer using strategies such as feature alignment or fine-tuning pretrained models. By contrast, in this approach we use a model's ability to transfer to a new context without such adjustments as a measure of the

generality of the conceptual framework learned by the model.

This general strategy can be implemented using the social media platform Reddit. On Reddit, users (N =234 million) self-organize into communities called subreddits. These subreddits sometimes reflect a general perspective (such as r/politics and r/philosophy), but more often reflect specific interests (such as r/modeltrains or r/badminton). This platform thus allows us to identify people having a certain mental perspective by virtue of their choosing to post to a particular subreddit. This platform also allows us to observe the same individuals across a range of contexts, thus allowing us to address whether the same perspective is implemented across contexts. If people apply a general framework across a range of situations, then a classifier that learns to distinguish general perspectives such as political orientation or religion should be able to identify those perspectives in people's posts from an unrelated context. Moreover, this generalization should be possible in two directions. The classifier trained on text from a particular perspective should generalize to text from a wide range of perspectives, and a classifier trained on text from a wide range of perspectives should generalize to text from a particular perspective. Lastly, this generalization should also apply over time. A conceptual framework is presumably a perspective that tends to remain relatively constant over time, and hence we should be able to use people's posts from an earlier point in time to predict their posts in the future. We tested these predictions in three lines of research. First, we looked at the transferrability of a religious perspective. Second, we looked at the generalizability of mental disorders. Lastly, we looked at whether or not these generalizations can be extended into the future.

Study 1: Religion

If people's online posts reveal something about their conceptual framework, then a model should be able to use the text of these posts to predict a person's conceptual framework. In Study 1 we asked this question by training a machine learning model to use people's Reddit posts to predict which religious subreddit the individual posts on. We evaluated this model in two separate contexts. First, we evaluated contexts where people explicitly write about religion (such as on the subreddit r/christianity). Second, we evaluated contexts where people write about more everyday topics (such as r/movies or r/travel). Finally, to evaluate whether the model learns a general representation across contexts, we asked whether a model trained in one context (e.g. writing on r/christianity) can make accurate predictions about people's religion in another context (e.g. writing on r/movies).

Methods

Data Acquisition. We acquired two separate datasets of Reddit posts: a *religious-context* set of posts submitted to subreddits about religion, and a *non-religious-context* set of posts submitted to other subreddits not about religion.

Religious-context dataset. We used a Reddit API (reddit.com/dev/api) to download 5 years of submissions (2012-2017) to 5 religious subreddits (r/Atheism, r/Buddhism, r/Christianity, r/Hinduism, r/Islam). 686,453 posts were obtained (range 6,773-418,229 posts/subreddit). We randomly undersampled these posts to create a balanced dataset of 33,865 posts, 6,773 posts/subreddit. This dataset was randomly divided into a *training set* (27,092 posts, 80%) for model training and a *test set* of (6,773 posts, 20%) for model evaluation.

Non-religious context dataset. For each user in the religious-context dataset, we downloaded all of the user's posts to all subreddits, excluding the religious subreddits above. Users who posted to more than one religious subreddit were excluded. We then created a dataset associating all of the posts for each individual (concatenated into a single post) with the religious subreddit that the individual had posted to. This dataset consisted of 127,698 users and was randomly undersampled to create a balanced dataset of 4,810 users (962 users/subreddit). The data was randomly divided into a training set (3,848 posts, 80%) and a test set (962 posts, 20%).

Data Preprocesing. Data preprocessing involved two steps. First, we removed names of religions from the posts, removing the words *atheist, atheism, buddhist, buddhism, christian, christianity, islam, muslim, islamic, hindu,* and *hinduism*, words beginning with *atheis, christian, buddh, islam,* or *hindu*, and non-ASCII characters. Second, we converted posts into machine-readable vectors. We used the vocabulary from each dataset separately to create *tf-idf* transformed 1gram vectors of length 257,559 (religious-context) and length 489,339 (non-religious context). Each element in the vector represents the frequency of writing a particular word, and the *tf-idf* transformation down-weights frequent but uninformative words such as *and* or *the*.

Machine Learning Model. We trained a machine learning model to use the words in people's posts to predict the subreddits they submitted to. We trained an L2-penalized logistic regression model to predict the probability that a post was submitted to each subreddit using its 1gram vector as input. The model was implemented in the python library scikit-learn with default regularization strength C=1 (for multi-class classification, scikit-learn fits a series of 5 one-vs-rest classifiers e.g. r/atheism vs. others). Model accuracy was calculated using the F score: $F = 2 * (\text{precision}*\text{recall}) / (\text{precision}+\text{recall})$. In this and future studies, p values were calculated by creating a null distribution based on 10,000 random draws given the observed sample size.

Results and Discussion

Religious context. If people's language reveals information about their religion, then when people explicitly write about religion, this language should be predictive of which

religion the individual is writing about. As would be expected for a model with this number of parameters, the model learned to predict religion in training data with high accuracy (81%, where chance = 25%), as also reflected in a high *F* score, 0.81. A more interesting test is whether the model generalizes to previously unseen test data without further training, which would suggest the model has learned to fit signal rather than noise in the training data. As shown in Table 1, the model had strong performance on held-out test data, $F = 0.77$, accuracy = 77% where chance = 0.20. The classifier performed best for Buddhism ($F = 0.82$) and Hinduism ($F = 0.80$) and worst for Atheism ($F = 0.62$), and performance was above chance for all religions as revealed by permutation testing ($p < 0.05$ for all classes). In future comparisons we report only this stricter test of performance based on held-out test data (not performance on the training data).

Non-religious context. If people's language reveals information about their religion, then when they talk about everyday topics such as movies or travel, this talk may still be revealing of which religious subreddit the individual also belongs to. People's everyday language was revealing of their religion, as revealed by moderate performance of the classifier on held-out test data, $F = 0.43$, accuracy = 43% where chance = 0.20. The classifier performed best for Atheism ($F = 0.50$) and worst for Hinduism ($F = 0.35$), and performance was above chance for all religions ($p < 0.05$ by permutation testing).

Transfer learning. If people's language reveals a conceptual framework that is consistent across contexts, then learning information about how language relates to religion in one context should also provide information about how language relates to religion in another, previously unseen, context. To test this, we asked whether the machine learning classifiers trained in either the religious or non-religious context could generalize, with no further training, to the other context. Both classifiers generalized moderately well, although generalization was strongest for the classifier trained on the non-religious context. The classifier trained on the religious context transferred well to the non-religious context ($F = 0.41$, accuracy = 41%, compared to $F = 0.43$ for a classifier trained and tested on the non-religious context). In addition, the classifier trained on the non-religious context transferred well to the religious context ($F = 0.43$, accuracy = 47%, compared to $F = 0.77$ for a classifier trained and tested on the religious context). The accuracies differed significantly from chance ($p < 0.05$ by permutation testing). These results suggest that the classifiers were able to pick up on people's conceptual frameworks, although they also learned to categorize the people on the basis of context specific features.

Discussion. The results showed that people's Reddit posts are diagnostic of an element of their conceptual framework: religion. When people explicitly talked about religion, a machine learning classifier was able to use the text of these posts to predict which religion people were writing about with high accuracy. When people talked about everyday topics such as movies or travel, their religious subreddit affiliation could still be identified with moderate accuracy. Note that it is possible that some of the everyday subreddits could have included religious content (for example we exclude r/hinduism but not r/india), but because these subreddits covered all of reddit, most were likely non-religious in topic. The conceptual framework learned was quite general, as evidenced by the ability of both classifiers to make accurate predictions in a novel context.

Table 1: Religion Classification Performance

	Train Context	
Test Context	Religious	Non-Religious
Religious	0.77	0.43
Non-Religious	0.41	0.43

Notes: Classification reported as *F* score, chance = 0.20. All values significantly differ from chance at $p < 0.05$.

Study 2: Clinical Psychological Disorders

Study 1 revealed that people's Reddit posts predict one kind of conceptual framework: religion. Study 2 investigated another kind of conceptual framework, the perspective people bring to experience from having a mental disorder. This was accomplished using Reddit posts focusing on clinical psychological disorders. There are many subreddits for clinical psychological disorders such as r/Depression and r/Anxiety. We downloaded people's posts to 4 common clinical psychological subreddits, and used the same methods as Study 1 to train a classifier to use these posts to predict individuals' membership to subreddits focusing on clinical psychological disorders. If people's Reddit posts reveal a conceptual framework, then a machine learning classifier trained on their Reddit posts should perform above chance in classifying membership to subreddits focusing on clinical psychological disorders.

Methods

Data Acquisition. We acquired two datasets of Reddit posts: a *clinical-context* dataset and a *non-clinical-context* dataset.

Clinical-context dataset. We used the same methods as Study 1 to download 5 years of posts (2012-2017) to 4 clinical subreddits: r/ADHD, a/Anxiety, r/Bipolar, r/Depression. 515,378 posts were acquired, which were undersampled to create a balanced dataset of 224,036 posts (56,009 posts/disorder), randomly split into a training set (179,228 posts, 80%) and a testing set (44,808 posts, 20%).

Non-clinical context dataset. For each user in the clinical-context dataset, we used the same methods as Study 1 to download all of the user's posts to non-clinical subreddits. We acquired posts for 121,722 users, randomly undersampled to create a balanced dataset of 24,436 users

(6,109 users/disorder) and randomly split into a training set (19,548 users, 80%) and a testing set (4,888 users, 20%).

Data Preprocessing and Machine Learning Model. Data preprocessing and the machine learning model were identical to Study 1 with 2 small changes. In preprocessing we removed the words *anxiety, anxious, depression, depressed, bipolar, adhd*, well as words beginning with *anx*, *depr*, *bipol*, and *add*, and non-ASCII characters. Also, new 1gram vectors were generated based on each dataset.

Results

Clinical Context. As with religion, people's posts on clinical subreddits were highly diagnostic of the clinical subreddit they was submitted to. As shown in Table 2, the classifier achieved $F = 0.77$, accuracy = 77% on held-out test data where chance = 0.25. Performance was highest for ADHD ($F = 0.84$) and lowest for depression ($F = 0.74$), and performance was above chance for all classes ($p < 0.05$ by permutation testing).

Non-Clinical Context. As with religion, people's posts in non-clinical contexts were diagnostic of the clinical subreddit they had also submitted to. As shown in Table 2, the classifier achieved $F = 0.38$, accuracy = 38% on held-out test data where chance = 0.25. Performance was highest for depression ($F = 0.44$) and lowest for anxiety ($F = 0.32$) and performance was above chance for all classes ($p < 0.05$ by permutation testing).

Transfer Learning. As in Study 1, we tested the ability of the clinical classifier to learn a general conceptual framework by using the classifiers trained on the clinical and non-clinical context to predict, with no further training, data from the other dataset. As shown in Table 2, both classifiers generalized well although generalization was strongest for the classifier trained in the clinical context. The classifier trained on the clinical context transferred well to the non-clinical context ($F = 0.37$, accuracy = 38%, compared to F=0.38 for a classifier trained and tested in the non-clinical context). In addition, the classifier trained on the non-clinical context transferred well to the clinical context ($F = 0.55$, accuracy = 56%, compared to $F = 0.77$ for a classifier trained and tested in the clinical context). The accuracies differed significantly from chance ($p < 0.05$ by permutation testing). As with religion, the results show that the classifiers learned a conceptual framework which was largely invariant across contexts.

Discussion. The results of Study 2 showed that people's Reddit posts are diagnostic not only of religion, but also of information about clinical psychological disorders. As in Study 1, people's explicit talk about clinical disorders was highly predictive of the clinical disorder being discussed. The classifier was also moderately accurate in using people's everyday language to identify which clinical subreddit an individual had submitted to, suggesting the models are able to learn a mindset that crosses over contexts. Again, we note that it is possible that some of these everyday subreddits overlapped with clinical topics (for example we exclude r/adhd but not r/psychiatry), but most were likely non-clinical. Transfer learning tests showed that both classifiers learned representations that were moderately predictive in another context, suggesting the model learns a framework that is consistent across contexts. Finally, the accuracies in all of these comparisons were similar to the model accuracies for predicting religion in Study 1, again suggesting that Reddit posts contain information about a broad conceptual framework.

Table 2: Clinical Classification Performance

	Train Context	
Test Context	Clinical	Non-Clinical
Clinical	0.77	0.55
Non-Clinical	0.37	0.38

Notes: Classification reported as F score, chance = 0.25. All values significantly differ from chance at $p < 0.05$.

Study 3: Predicting the Future

Studies 1-2 revealed that people's language was predictive their conceptual framework, even when this language was about everyday topics. Interestingly, these studies relied on posts submitted on any date, including posts from before an individual ever joined a clinical or religious subreddit. The success of these classifiers suggests it may be possible to predict a person's future conceptual framework. To assess this possibility, we re-trained the religion and clinical classifiers using only posts from before an individual joined a religious or clinical subreddit. We asked whether this past language was predictive of which religious or clinical subreddit the individual joined in the future. If people's past language predicts their future conceptual framework, then both classifiers should perform above chance. Additionally, the performance of these classifiers can be compared to the classifiers from studies 1-2 to assess how predictive the past is relative to all of an individual's posts.

Methods

Data Acquisition. Two datasets were acquired: the *past-religion* and *past-clinical* datasets. These datasets consisted of all posts in the *non-religious-context* and *non-clinical-context* datasets that were submitted before that user every posted to a religious or clinical subreddit. The past-religion dataset, after undersampling, consisted of 2,630 users, split randomly into a training set (2,104 users, 80%) and a testing set (526 users, 20%). The past-clinical dataset, after undersampling, consisted of 18,040 users, split randomly into a training set (14,432 users, 80%) and a testing set (3,608 users, 20%). Data preprocessing was identical to studies 1-2.

Data Preprocessing and Machine Learning Model. Methods were identical to Study 2 except that new 1gram vectors were generated for each dataset.

Results and Discussion

Religion. As expected, people's past posts were predictive of their future conceptual frameworks. People's past posts from non-religious contexts predicted which religious subreddit they later posted to (F = 0.36, accuracy = 36% on held-out test data where chance = 0.20). The classifier performed best for Atheism (F = 0.41) and worst for Buddhism (F = 0.30), and performance was above chance for all classes ($p < 0.05$ by permutation testing). Strikingly, performance of the classifier trained to predict the future was almost as high the classifier trained in Study 1 to predict the present (F = 0.36 vs. 0.43).

Clinical Disorders. People's past posts from non-clinical contexts predicted which clinical subreddit they later posted to (F = 0.36, accuracy = 36% on held-out test data where chance = 0.25). The classifier performed best for depression and bipolar disorder (both F = 0.39) and worst for anxiety (F = 0.29), and performance was above chance for all classes ($p < 0.05$ by permutation testing). Strikingly, performance of the classifier trained to predict the future was almost as high as the classifier trained in Study 2 to predict the present (F = 0.36 vs. 0.38).

Discussion. The results of Study 3 show that people's Reddit posts are not only predictive of their current conceptual framework, but also of their future conceptual framework. Classifiers trained to use people's past posts to predict their affiliation with religious or clinical subreddits performed almost as well as classifiers based on past and future posts. A limitation of Study 3 is that the date a user subscribes to a subreddits is an imperfect indicator of mental phenotype. For example, an individual may have depression before choosing to post to the r/depression subreddit. Nevertheless, the similar performance of models trained to predict the future as models trained to predict the present suggests that a large component of what the models are learning is a consistent framework over time. We note also that some of the conceptual frameworks studied, particularly one's religious affiliation, do not always change over time. In the case of religion, one possibility is that our model learns relative changes in the intensity of the conceptual framework (for example, joining a religious subreddit may be a sign of becoming more religious over time).

Study 4: Representations

When a classifier learns a conceptual framework, it in effect learns a set of words that predict that framework. These words are the model's representation of the framework. An examination of these representations may prove valuable in at least two ways. First, an examination of the words can offer evidence for the hypothesis that the classifiers are, in fact, learning something about the conceptual frameworks. If the classifier is working, then the words should form coherent semantic clusters with clear associations to the conceptual framework. Second, assuming the classifiers form coherent semantic clusters, it may then be possible to use these clusters to gain some insight into the nature of the conceptual frameworks. It may be possible, for example, to gain insight into what is experienced by someone who is depressed or what is emphasized in a Christian world view. These two aims were pursued in the current study.

Methods

Classifier and Feature Selection. The solutions learned by classifiers in Study 1 (the religious-context and non-religious-context classifiers) and Study 2 (the clinical-context and non-clinical-context classifiers) were analyzed. For each classifier we selected the 100 words with the highest regression weights.

Clustering Analysis. For each set of 100 words, we performed a cluster analysis. This analysis had three steps. First, the semantics of each word was specified using pre-learned vectors trained on part of the Google News dataset, which is based on about 100 billion words and contains approximately 3 million words and phrases (https://code.google.com/archive/p/word2vec/). Training used the Word2Vec learning procedure (Mikolov et al, 2013). The second step was to reduce the dimensionality of each set 100 vectors down from 300 to 2 dimensions using the t-SNE algorithm (Maaten & Hinton, 2008). t-SNE was preferred to PCA because it prioritizes local over global spatial information, which is especially important for clustering analysis. In the final step, clusters were identified using the k-means++ cluster algorithm, which tries to separate elements into groups by minimizing within-cluster sum-of-squares. The number of clusters was determined by running the algorithm for different numbers of k and choosing the k that maximized the Silhouette Coefficient (Rousseeuw, 1987).

Results

As expected, the top 100 words for the different religions and mental illness fell into coherent semantic clusters. Example clusters for each conceptual framework are listed in Table 3. Also as expected, the content of the clusters provides some insight into the nature of the conceptual frameworks. For example, those with anxiety tended to mention being nervous and their breath. Those with depression mentioned feelings of despair and meaninglessness. The findings from these clusters are not necessarily surprising. Importantly, they were derived automatically and allowed for good classification, even

when the text concerned topics unrelated to religion and mental disorders.

Table 3: Word Clusters associated with Conceptual Frameworks

Framework	Words
Christianity	angels, demons, heaven, lewis, resurrection, salvation, sin, sinful, sins, soul
	communion, gospel, holy, kingdom, sermon
Buddhism	elightened, philosophy, precepts, teachings, tradition, truths, wisdom
	attachment, mind, nature, path, realms, rebirth, sn
Hinduism	sanskrit, scriptures, shaivism, upanishads, vedas, vedic
	bhajan, bhakti, chalisa, gita, kirtan, moksha, namaste
Islam	fasting, hajj, halal, iftar, ramadan, ramadhan
	alaikum, alaykum, duas, fajr, haram, mecca, pbuh,
Atheism	argument, evolution, exist, ignorance, intelligent, logic, santorum, science
	actually, crazy, fun, just, ridiculous, scary, stupid
Adhd	attention, concentration, productive, productivity, task
	hyper, hyperfocus, hyperfocused, impulse, sensory, stimulating, stimulation
Anxiety	afraid, freaked, freaking, nervous, panic, panicked, panicking, panicky, scared, terrified
	breath, breathe, breathing, chest, heart, shaking, stomach
Bipolar	episode, episodes, highs, lows, psyc, swing, tracking
	fearless, grandiose, mania, manic, mood, moods, rage
Depression	darkness, despair, emptiness, hopelessness, loneliness, worthlessness
	meaningless, miserable, pathetic, pointless, shittier, shitty, throwaway, ugly, worthless

Notes: Example clusters of words associated with religions and mental disorders from Study 4.

General Discussion

The results support the existence of a general mental framework that people apply across different contexts. In a series of studies, machine learning classifiers were trained to use the text of people's posts on Reddit to predict two aspects of their conceptual framework: religion and clinical psychological disorders. As in prior work, the classifiers could use people's language to identify their conceptual framework, both when individuals were explicitly talking about religion and clinical disorders and when they were talking about other everyday topics. Extending prior work, a classifier trained in one context showed moderate to high generalization to another context with no additional training, suggesting the representations learned by the model are indeed a stable conceptual framework. Finally, people's past language was almost as predictive of their future conceptual framework as their past and future language combined. This ability to predict the future suggests that the conceptual framework learned by the model is quite stable over time.

The results also have implications for the automated identification of clinical psychological disorders. An open question is whether automated methods capture variance that would not be captured by existing clinical methods, for example by identifying previously undiagnosed cases (Guntuku et al, 2017). The results of Study 3 suggest that one way classifiers add to clinical diagnosis is by identifying signs that an individual may develop a disorder in the future. Thus, automated methods may be especially useful for predictively identifying whether an individual is likely to develop a disorder in the future.

The results have a few limitations. First, posting on a particular subreddit is an imperfect indicator of an individual's cognitive framework. For example, an individual can post on r/depression without a clinical diagnosis of depression. However, the classifiers learned features with high face validity, suggesting that people's talk in these contexts reflects what we think of as a cognitive framework. Second, the time an individual joins a forum is an imperfect indicator of their cognitive framework. For example, an individual may have depression before they join r/depression. However, the strong performance of the classifier trained on past posts alone suggests that a large portion of people's cognitive framework is stable over time.

Overall, the results suggest that people have a general mental framework that they apply across contexts, and that these frameworks can be identified using machine learning methods.

References

Guntuku, S., Yaden, D., Kern, M., Ungar, L., & Eichstaedt, J. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43-49.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, *110*(15), 5802-5805.

Liu, L., Preotiuc-Pietro, D., Samani, Z. Moghaddam, M. E., & Ungar, L. (2016, May). Analyzing Personality through Social Media Profile Picture Choice. In *ICWSM* (pp. 211-220).

Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*, 2579-2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Schwartz, H., Eichstaedt, J., Kern, M., … Ungar, L. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one, 8*(9), e73791.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS, 112*(4), 1036-1040.